Text mining: de volgende stap in zoektechnologie : vinden, zonder precies te weten wat men zoekt of vinden wat er niet lijkt te zijn

Text mining: de volgende stap in zoektechnologie

Colofon

Ontwerp en print: Océ Business Services, Maastricht

ISBN: 978-90-5681-306-2
NUR: 740

Text mining: de volgende stap in zoektechnologie

Vinden, zonder precies te weten wat men zoekt of vinden wat er niet lijkt te zijn

Inaugurele rede uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar aan de afdeling Knowledge Engineering van de faculteit Humanities and Sciences aan de Universiteit van Maastricht

Maastricht, 23 januari 2009

Dr. ir. Jan C. Scholtes

1 Inhoudsopgave

2 Saluut en Inleiding

Mijnheer de rector magnificus, hooggeleerde collega's, en geachte andere aanwezigen, allemaal van harte welkom en dank voor het aanwezig zijn bij deze inaugurele rede, waarmee ik mijn ambt als bijzonder hoogleraar in de *text mining* aan de Universiteit van Maastricht zal aanvaarden.

Ik ben vereerd en dankbaar tegelijk dat wij hier vanmiddag aanwezig zijn en dat ik in de gelegenheid ben om een klein uur te kunnen uitweiden over het uitermate interessante onderwerp *text mining*. Zeker omdat een groot aantal van de hier aanwezigen, het meestal al na paar minuten voor gezien houden, als ik ze probeer uit te leggen wat mij de hele dag bezighoudt. ☺

Binnen het vakgebied *text mining*, soms ook wel *text analytics* genoemd, komen een aantal interessante technologieën samen zoals computers, informatica, computationele linguïstiek, cognitie, patroonherkenning, statistiek, geavanceerde wiskundige technieken, artificiële intelligentie, visualisatie en niet te vergeten *information-retrieval*. Allemaal onderwerpen waar ik me de afgelopen vijfentwintig jaar met veel plezier en interesse in heb verdiept.

De komende jaren hoop ik hier samen met de studenten en collega's van de Universiteit van Maastricht verder aan te werken zodat we over een paar jaar kunnen zeggen dat we vooruitgang hebben geboekt.

Vanmiddag zie ik het als mijn taak om u allen mee te nemen voor een korte rondleiding door mijn vakgebied en een blik in de toekomst. Eerst wordt duidelijk gemaakt wanneer *text mining* technologie relevant is. Dit zal gebeuren aan de hand van een aantal zoekproblemen. Hierna wordt dieper ingegaan op de verschillende technologieën die voor het vakgebied van belang zijn en er zullen diverse voorbeelden geven worden van succesvolle toepassingen van *text mining*. Ook wordt er kort ingegaan op deelgebieden binnen de *text mining* waar meer onderzoek gewenst is.

De informatie explosie van de laatste decennia zal namelijk in hetzelfde tempo doorgaan. U bent ongetwijfeld allemaal bekend met de bekende wetten van Moore, één van de oprichters van Intel en mede-uitvinder van de computerchip: volgens Moore verdubbelt iedere achttien maanden de reken- en opslag capaciteit van een computer, deze wet geldt al sinds

de jaren vijftig van de vorige eeuw. Door dit exponentiële gedrag zijn wij iedere achttien maanden in staat twee keer zoveel informatie te verwerken of op te slaan. Dit resulteert in een steeds grotere *information-overload* en in het steeds moeilijker terugvinden van informatie aan de ene kant, maar tevens in mogelijkheden voor diverse nieuwe computertechnieken die ons helpen deze berg aan informatie te controleren aan de andere kant. Deze nieuwe technieken moeten natuurlijk wel ontwikkeld worden.

Text mining technieken zullen de komende jaren een essentiële rol zal spelen in dit zich continue voortschrijdende proces.

3 Wat is Text Mining

Het vakgebied van *data mining* is bekender dan dat van *text mining*. Een goed voorbeeld van *data mining* is het analyseren van transactie gegevens die in relationele databases zitten. Denk aan creditcard betalingen of pin-transacties. Aan dergelijke transacties kan men diverse aanvullende kenmerken meegeven: datum, locatie, leeftijd van creditcard houder, salaris, etc. Met behulp van de combinatie van deze gegevens kunnen dan patronen van interesse of gedrag bepaald worden.

Echter, meer dan 90% van alle informatie is ongestructureerde informatie, en zowel het percentage als de absolute hoeveelheid ongestructureerde informatie groeien iedere dag. Slechts een beperkte hoeveelheid informatie is opgeslagen in een gestructureerd formaat in een database. De meeste informatie waar we dagelijks mee werken staat in tekst documenten, e-mails, of in multimediale (spraak, video, en foto's) bestanden. Daarin zoeken of analyses maken met database- of *data mining* technieken is onmogelijk. Deze werken namelijk alleen op gestructureerde informatie.

Het is makkelijker om gestructureerde informatie te doorzoeken, te beheren, te organiseren, te delen en er rapportages mee te maken. Niet alleen voor mensen, maar ook voor computers. Vandaar de wens om ongestructureerde informatie te structureren waarna zowel mensen als computers er beter mee om kunnen gaan èn omdat we dan ook ons bekende technieken en methodieken kunnen gebruiken.

In het midden van de jaren tachtig van de vorige eeuw werd text mining voor het eerst toegepast, dit waren vooral handmatige technieken. Al snel bleken deze handmatige technieken te arbeidsintensief en daardoor

te kostbaar te zijn. Ook duurde het veel te lang om de almaar groeiende hoeveelheden informatie met de hand te structureren. In de loop der jaren werd men steeds succesvoller bij het automatiseren van deze processen. Vooral de laatste tien jaar is er veel vooruitgang geboekt.

Tegenwoordig richt het vakgebied van de *text mining* zich vooral op het ontwikkelen van diverse geavanceerde wiskundige-, statistische-, taalkundige- en patroonherkenning technieken waarmee het mogelijk is om ongestructureerde informatie automatisch te analyseren alsmede om hoge kwaliteit en relevante gegevens te extraheren en de tekst in zijn geheel daardoor beter doorzoekbaar te maken.

Hoge kwaliteit refereert hier in het bijzonder aan de combinatie van relevantie (oftewel: de speld in de hooiberg vinden) en het verkrijgen van nieuwe interessante inzichten.

Een tekst document bestaat uit karakters, die samen woorden vormen, welke gecombineerd kunnen worden tot termen. Dit zijn allemaal syntactische eigenschappen die samen bepaalde categorieën, concepten, betekenissen of bedoelingen representeren. *Text mining* wil al deze informatie kunnen herkennen, extraheren en gebruiken.

Met behulp van *text mining* technieken kunnen we in plaats van zoeken op woorden, zoeken op taalkundige patronen van woorden, dit is dus zoeken op een hoger niveau!

Mede door de voortgaande globalisering, is er ook veel interesse voor meertalige *text mining*: het verkrijgen van inzichten over meertalige collecties. De recente beschikbaarheid van hoge kwaliteit machinale vertaalsystemen is in die context een belangrijke aanwinst. Meertalige *text mining* is complexer dan het lijkt, want naast de verschillen in karaktersets en woorden, maakt *text mining* ook intensief gebruik van zowel statistische als taalkundige (zoals vervoegingen, grammatica, betekenis, en bedoeling) eigenschappen van een taal.

Dagelijks maken wij al meer gebruik van *text mining* en andere eerder genoemde technieken dan u denkt, vaak onbewust. Een voorbeeld: als u op internet zoekt met een zoekmachine, dan kan het zo zijn dat u op maat gemaakte advertenties krijgt gepresenteerd als u een bepaald artikel leest. Dit zijn bijvoorbeeld advertenties die aansluiten bij de tekst

in het artikel of als u van bepaalde gratis email diensten gebruik maakt; er worden dan aan de hand van de woorden die in de tekst van het email bericht gebruikt worden specifieke advertenties getoond. Hiervoor worden *text mining* techieken gebruikt.

Text mining gaat dus verder dan dat de computer weet waar u zich bevindt, wat uw interesse is of wat uw leeftijd is. Het kan zelfs zo zijn dat via informatie in een sociaal netwerk gekeken wordt welke informatie vergelijkbare personen interessant vinden. Dit is allemaal nog gestructureerde informatie. Bij *text mining* gaat het om het analyseren van ongestructureerde informatie en daar relevante patronen en kenmerken uithalen.

Vervolgens kan men met die patronen en kenmerken beter zoeken, dieper data analyseren en sneller inzichten krijgen die anders vaak verborgen blijven.

Uitgaande van deze basis principes zijn er vele toepassingsgebieden, maar de belangrijkste bevinden zich op het gebied van "zoeken en vinden van informatie". Hier zullen we ons vandaag verder op concentreren.

4 Zoeken met Computers in Ongestructureerde Informatie

Wat gebeurt er precies als men met een computerprogramma zoekt in ongestructureerde tekst? Ik zal dit kort toelichten: Computers zijn digitale apparaten met beperkte mogelijkheden. Computers kunnen het beste omgaan met getallen, en als het echt snel moet zijn, dan gehele getallen in het bijzonder, ook wel *integers* genoemd. Mensen zijn analoog, onze menselijke taal is analoog, vol met inconsistenties, ruis, fouten en uitzonderingen. Als we iets zoeken, dan denken we vaak in concepten, betekenissen en bedoelingen, allemaal zaken waar een computer niet direct mee om kan gaan.

Voor men computers op een computationeel efficiënte manier kunt laten zoeken in grote hoeveelheden tekst, zal eerst het probleem vertaald moeten worden naar een getalmatig probleem waar een computer mee om kan gaan. Dit leidt tot hoogdimensionale ruimtes van heel veel getallen waar we dan getallen, die zoektermen representeren, vergelijken met getallen die documenten en informatie representeren. Dit is in de basis waar ons vakgebied zich mee bezig houdt: hoe kunnen we informatie zoals wij mensen die verwerken, vertalen naar informatie

die een computer kan verwerken en de uitkomst daarvan dan weer vertalen naar iets wat wij mensen begrijpen.

Deze technologie bestaat al sinds de jaren zestig van de vorige eeuw. Eén van de eerste wetenschappers die zich hiermee bezighield was Gerald Salton. Samen met anderen maakte hij één van de eerste tekst-zoekmachines. Men sloeg het voorkomen van ieder woord in een document op in een trefwoorden index. Zoeken vond dan plaats op de index, vergelijkbaar met een index achter in een boek, maar dan op veel meer woorden en vele malen sneller. Technieken als hashing en b-trees maakten het mogelijk om snel en efficiënt een lijst te krijgen van alle documenten waarin een bepaald woord of een Booleaanse (*AND, OR* en *NOT* operatoren) combinatie van woorden voorkwam.

Documenten en zoekvragen werden vertaald naar vectoren en vergeleken via de Cosinus afstand tussen beiden: hoe kleiner de Cosinus afstand, hoe meer de zoekvraag en het document overeenkwamen. Dit was een effectieve manier om de relevantie van documenten te bepalen gegeven een bepaalde zoekvraag. Dit werd het *vector space* model genoemd en wordt tot op de dag van vandaag door sommige programma's nog steeds gebruikt.

Later werden diverse andere manieren van zoeken en relevantie onderzocht. Er zijn tientallen zoektechnieken met welklinkende namen als: (*directed en non-directed)-proximity, fuzzy, wildcards, quorum, semantical, taxonomies, conceptual, etc.* Bekende voorbeelden van relevantie bepalingen zijn *term-based frequency ranking, het page-rank algoritme (populariteitsbeginsel), en probabilistic ranking (Bayes classifiers)*

Salton's eerste grote publicatie was in 1968, nu éénenveertig jaar geleden. Zijn alle problemen gerelateerd aan zoeken en vinden dan nog niet opgelost, zult u zich afvragen?

Het antwoord is *nee*. Omdat er tegenwoordig zoveel informatie digitaal beschikbaar is en omdat het tegenwoordig noodzakelijk is om vaak direct (pro-actief) te kunnen reageren op wat er gebeurd, zijn nieuwe technieken nodig om bij te kunnen blijven met de almaar groeiende hoeveelheid ongestructureerde informatie. Daarnaast zijn er verschillende redenen en doelen waarom iemand veel data wil doorzoeken en deze verschillen resulteren in de noodzaak tot verschillende manieren van aanpak.

5 Text Mining in Relatie tot "Zoeken & Vinden"

De titel van deze voordracht luidt: "*text mining*: de volgende stap in zoektechnologie"; met als ondertitel: "Vinden zonder precies te weten wat men zoekt", en "vinden wat er niet lijkt te zijn". Hoe doet men dat? Wie willen dat? Of in andere woorden: wat is de maatschappelijke dan wel de wetenschappelijke relevantie hiervan.

Zo werd mij ook gevraagd tijdens het sollicitatie proces voor dit hoogleraarschap: "We hebben Google toch, dus wat hebben we nog meer nodig?". "Een heel goede vraag", was mijn reactie, "want dit is precies zoals veel mensen er over denken". Helaas is het zoekprobleem nog niet opgelost en Google geeft niet het volledige antwoord op al uw vragen. Als ik u hiervan in de komende vijfenveertig minuten kan overtuigen, dan ben ik alvast in dat deel van mijn missie geslaagd!

Men zou de vragen die ik net gesteld heb ook anders kunnen formuleren:

"Wil men alleen het beste vinden of wil men alles vinden" of "Wil men vinden wat en wie niet gevonden wil worden".

5.1 Alles vinden

Dan komen we al dichter bij de essentie van het probleem. Internet zoekmachines geven alleen de beste antwoorden, of de meest populaire antwoorden. Fraude onderzoekers of juristen willen niet alleen de *beste* documenten, ze willen *alle mogelijk* relevante documenten.

Verder doet bij een internet zoekmachine iedereen zijn best om boven in de resultatenlijst te staan: zoekmachine optimalisatie is een wetenschap op zich geworden. Criminelen en fraudeurs willen niet boven in de resultaatlijst van een zoekmachine staan. Ze proberen juist te verbergen wat ze doen.

5.2 Vinden wie of wat niet gevonden wil worden

Hoe doen ze dat: ze gebruiken synoniemen, code namen, vaak zijn dit veel voorkomende woorden die zo vaak voorkomen dat er nooit op gezocht kan worden zonder miljoenen treffers te krijgen. Om toch dit soort relevante informatie te kunnen vinden kan text mining een uitkomst bieden.

5.3 Vinden terwijl men niet precies weet wat men zoekt

Fraude onderzoekers hebben ook een ander gemeenschappelijk probleem: ze weten aan het begin van een onderzoek vaak niet precies waar ze op moeten zoeken. Ze kennen de synoniemen en code namen niet of ze weten niet precies op welke bedrijven, personen, rekeningnummers, bedragen, ze moeten zoeken. Met *text mining* is het mogelijk om al dit soort entiteiten of eigennamen aan de hand van hun taalkundige rol te identificeren en ze vervolgens te classificeren en op een gestructureerde manier aan een gebruiker te presenteren. Het is dan heel eenvoudig om de voorkomende bedrijven of individuen verder te onderzoeken.

Soms gaat het probleem van een onderzoeker nog een stapje verder: ze zoeken terwijl ze niet precies weten wat ze zoeken. Om de woorden en onderwerpen te vinden die van belang zijn voor het onderzoek kan men *text mining* gebruiken: de computer zoekt naar bepaalde patronen in de tekst: "wie betaalt wie wat", "wie praat met wie", etc. Met taaltechnologie en *text mining* kunnen dit soort patronen herkend worden, uit de tekst gehaald worden en aan een onderzoeker gepresenteerd worden. Die zal dan snel kunnen bepalen wat legitieme transacties zijn en wat opvallende transacties zijn.

Een voorbeeld: als de ABN-AMRO geld overmaakt naar de FORTIS dan is dat een normale transactie. Maar als "Grote Tinus" geldt overmaakt naar de Bahamas Enterprises Inc., dan is dat wellicht verdacht. Met *text mining* kunnen dit soort patronen dus geïdentificeerd worden en vervolgens kun men op de woorden in die patronen met normale zoektechnieken doorzoeken en de gegevens verder identificeren en analyseren.

Het verkrijgen van nieuwe inzichten wordt ook wel serendipiteit genoemd: serendipiteit (afgeleid van het Engelse woord *serendipity*: het vinden van iets onverwachts en bruikbaars terwijl men eigenlijk op zoek is naar iets totaal anders). *Text mining* kan heel goed toegepast worden voor het verkrijgen van dit soort nieuwe maar vaak noodzakelijke inzichten om verder te komen bij een groot onderzoek.

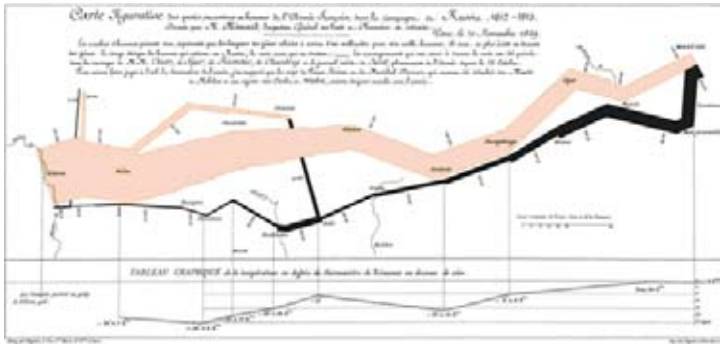
We kunnen dus zeggen dat *text mining* helpt bij het vinden van informatie middels patronen waarvan de waardes van de elementen van te voren niet exact bekend zijn. Vergelijkbaar met wiskundige functies waarbij de variabelen en de statistische distributie van de variabelen niet altijd

bekend zijn. Ook hier kan de essentie van het probleem gezien worden als een vertaalprobleem van menselijke taal naar de wiskunde. Hoe beter de vertaling, hoe beter de kwaliteit van de *text mining*.

5.4 Text mining en informatie visualisatie

Text mining wordt vaak in één zin genoemd met informatie visualisatie. Dit komt omdat visualisatie één van de technische mogelijkheden is, die mogelijk wordt nadat ongestructureerde informatie is gestructureerd.

Een voorbeeld van informatie visualisatie is de zogenoemde bewegingskaart van M. Minard uit 1869 die een inzicht geeft in Napoleon's mars naar Rusland. De breedte van de lijn geeft het aantal manschappen aan tijdens de campagne. Goed is te zien dat gedurende de heen en terugweg het aantal manschappen dramatisch afneemt.



Figuur 1: M. Minard (1869): Napoleon's expeditie naar Rusland (Bron: Tufte, Edward, R. (2001). The Visual Display of Quantitative Information, 2nd edition).

Deze kaart geeft sneller een beter inzicht dan rijen met getallen. Dat is kort samengevat de essentie van informatievisualisatie: een plaatje zegt vaak meer dan duizend woorden.

Om dit soort visualisaties te kunnen maken, moeten gegevens gestructureerd zijn. Dit is precies waar *text mining* technologie kan helpen: door ongestructureerde gegevens te structureren is het mogelijk om de data te visualiseren en sneller nieuwe inzichten te krijgen.

Een voorbeeld is de volgende tekst:

ZyLAB donates a full ZylMAGE archiving system to the Government of Rwanda

Amsterdam, The Netherlands, July 16th, 2001 -ZyLAB, the developer of document imaging and full-text retrieval software, has donated a full ZylMAGE filing system to the government of Rwanda.

"We have been working closely with the UN International Criminal Tribunal in Rwanda (ICTR) for the last 3 years now," said Jan Scholtes, CEO of ZyLAB Technologies BV. "Now the time has come for the Rwanda Attorney General's Office to prosecute the tens of thousands of perpetrators of the Rwanda genocide. They are faced with this long and difficult task and the ZyLAB system will be of tremendous assistance to them. Unfortunately, the Rwandans have scarce resources to procure advanced imaging and archiving systems to help them in this task, so we decided to donate them a full operational system."

"We greatly thank you for this generous gift," says The Honorable Gerald Gahima, the Rwandan Attorney General. "We possess an enormous evidence collection that will require scanning so we can more effectively process, search and archive the evidence collection."

A demonstration of the ZyLAB software was done for the Rwandans by David Akerson of the Criminal Justice Resource Center, an American-Canadian volunteer group: "The Rwandans were greatly impressed. They want and need this system as they currently have evidence sitting in folders that is difficult to search. This is one of the major delays in getting the 110,000 accused persons in custody to trial."

"My hope and belief is that ZylMAGE will enable Mr. Gahima's office to process, preserve and catalogue the Rwandan evidence collection, so that the significance and details of the genocide in Rwanda can be preserved," Scholtes concludes.

In deze tekst kunnen o.a. de volgende entiteiten en attributen gevonden worden:

Plaatsen	Amsterdam
Landen	The Netherlands, Rwanda
Personen	Jan Scholtes, Gerald Gahima, Mr. Gahima's, David Akerson, Scholtes
Functienamen	CEO, Rwandan Attorney General

Data	July 16th, 2001
Organisaties	UN International Criminal Tribunal in Rwanda (ICTR), Government of Rwanda, Rwanda Attorney General's Office, Criminal Justice Resource Center, American-Canadian volunteer group
Bedrijven	ZyLAB, ZyLAB Technologies BV
Producten	ZyIMAGE

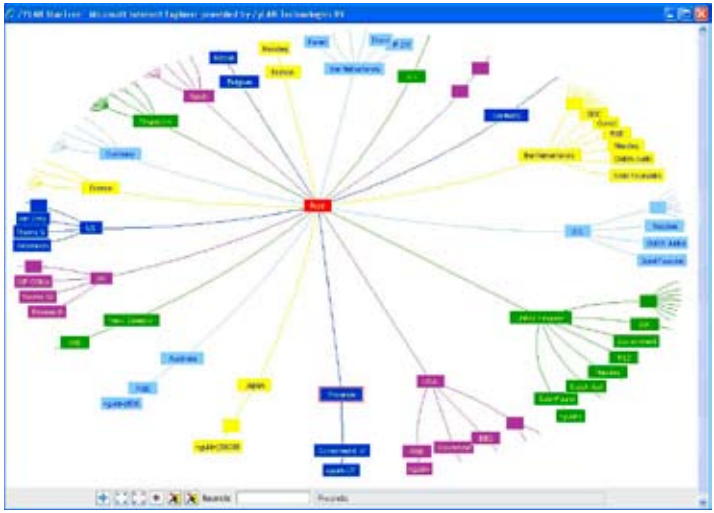
Stel nu dat men diverse documenten heeft met dit soort automatisch gevonden gestructureerde eigenschappen, dan kan men de documenten niet alleen in tabelvorm laten zien, maar ook bijvoorbeeld in een boomstructuur waarbij men de documenten eerst op de voorkomens per land en dan op de voorkomens per organisatie organiseert. Dit kan dan worden ingeladen in bijvoorbeeld een *Hyperbolic Tree* of in een zogenaamde *TreeMap*.

Beiden geven de mogelijkheid om op de delen van de boomstructuur waarin men geïnteresseerd is in te zoomen zonder het totaaloverzicht te verliezen.

Een goed voorbeeld van een weergave van een hyperbool (het principe waarop de *Hyperbolic Tree* is gebaseerd) is te vinden in het werk van onze Nederlandse M. C. Escher. Hierbij wordt een tweedimensionaal voorwerp op een bol gelegd en vervolgens zal het centrum automatisch inzoomen en de randen automatisch uitzoomen.

Figuur 2: M.C. Escher: Circle Limit IV 1960 woodcut in black and ocre, printed from 2 blocks (Bron: http://www.mcescher.com/).

Dit principe kan ook gebruikt worden om een boomstructuur dynamisch te visualiseren. In dat geval ziet de visualisatie er als volgt uit:

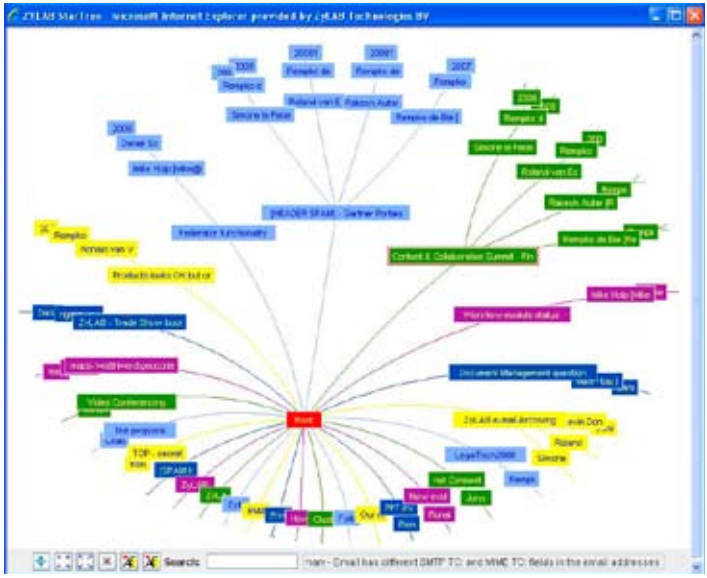


Figuur 3: Hyperbolic Tree visualisatie van een boomstructuur (bron: ZyLAB Technologies BV).

Figuur 4: TreeMap visualisatie van een boomstructuur. (bron: ZyLAB Technologies BV).

Een andere manier om een boomstructuur weer te geven is in een zogenaamde *TreeMap*, geïntroduceerd door Ben Shneiderman in 1992. Hierbij wordt een boomstructuur op een oppervlakte geprojecteerd en hoe meer bladeren er aan een bepaalde tak zitten, des te meer oppervlakte krijgt deze tak toegewezen. Op deze manier kan men snel zien waar de meeste entiteiten zich bevinden. Men kan per entiteit de grootte ook een bepaalde waarde laten weergeven. Bijvoorbeeld de grootte van een email of een bestand.

Dit soort visualisatie technieken zijn bij uitstek geschikt om grote collecties email snel inzichtelijk te maken. Hierbij kan naast de structuur die *text mining* technieken kunnen ontdekken, ook gebruik gemaakt worden van al aanwezige kenmerken zoals “Afzender”, “Ontvanger”, “Onderwerp”, “Datum”, etc. Hieronder zijn een aantal mogelijkheden van email visualisaties opgenomen.

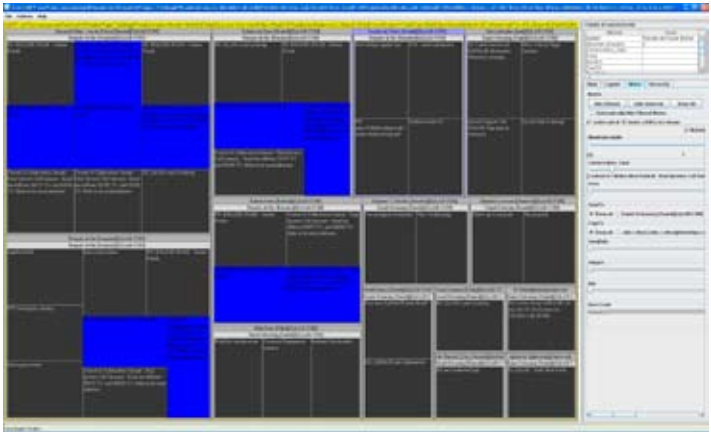


Figuur 5: Email visualisatie met een Hyperbolic Tree (bron: ZyLAB Technologies BV).

Met behulp van dit soort visualisatie technieken is het mogelijk om sneller en beter inzicht te krijgen in complexe dataverzamelingen, zeker als men te maken heeft met grote collecties ongestructureerde informatie die

door het gebruik van text mining snel en automatisch gestructureerd kunnen worden.

Figuur 6: Email visualisatie met een TreeMap (bron: ZyLAB Technologies BV).



Figuur 7: Email visualisatie met een TreeMap waarbij alle berichten uit één email conversatie gemarkeerd zijn met dezelfde kleur: direct is te zien wie bij een conversatie betrokken waren (bron: ZyLAB Technologies BV).

5.5 Andere voordelen van gestructureerde en geanalyseerde data

Naast de bovengenoemde visualisatie zijn er diverse andere toepassingen mogelijk als ongestructureerde data eenmaal gestructureerd is en van metagegevens is voorzien. Een aantal wordt hieronder opgesomd:

- Gegevens zijn makkelijker te organiseren in folders.
- Het is mogelijk om data te filteren op bepaalde metagegevens bij het zoeken of bekijken van data.
- Het is mogelijk om gegevens te vergelijken en te koppelen aan de hand van de metagegevens (vector vergelijkingen van metagegevens)
- Het is mogelijk om op basis van ieder van de kenmerken documenten te sorteren, te groeperen en te prioriteren.
- Gegevens kunnen worden geclusterd aan de hand van metagegevens.
- Aan de hand van de metagegevens kunnen duplicaat en bijna duplicaten worden herkend. Vervolgens kunnen deze of worden verwijderd of apart worden gezet.
- Het is mogelijk om taxonomieën af te leiden uit de metagegevens.
- Er kunnen zogenaamde *topic* analyses en *discourse* analyses gemaakt worden aan de hand van de metagegevens.
- Het is mogelijk om regelgebaseerde analyses op de metagegevens toe te passen.
- Het is mogelijk om door te zoeken op de metagegevens van reeds gevonden documenten.
- Diverse (statistische) rapportages kunnen gemaakt worden op basis van de metagegevens.
- Het is mogelijk om te zoeken naar relaties tussen metagegevens: bijvoorbeeld: "wie betaalt wie wat", waarbij de "wie" en de "wat" van te voren onbekend zijn.

Toepassingen van deze technieken zijn er op verschillende vakgebieden.

In de volgende sectie zal worden ingegaan op alledaagse toepassingen van *text mining* technologie. Daarna zullen we kort ingaan op de verschillende technieken die nodig zijn voor succesvolle *text mining* toepassingen.

6 Voorbeelden van Toepassingen van Text-Mining

6.1 Fraude, criminaliteitsopsporing, en inlichtingen analyses

Het moge duidelijk zijn dat er voor *text mining* grote toepassingsmogelijkheden zijn bij fraude- en criminaliteitsopsporing, inlichtingen analyses en vergelijkbare toepassingen. Het moet ook gezegd worden dat text mining zijn oorsprong vond in dit soort toepassingen en dat het in deze vakgebieden tegenwoordig zelfs onmogelijk is om succesvol en efficiënt te werken zonder *text mining* technologie.

Een mooi voorbeeld is hieronder te vinden. In de tekst zijn zowel individuele entiteiten herkend en weergegeven in een bepaalde kleur (eerste vijf van *Person* tot *Weapon*), als patronen van entiteiten die herkend zijn (laatste zeven). Vooral de laatste zeven herkende patronen zijn erg interessant. In deze gevallen was het mogelijk om met behulp van *text mining* bepaalde interessante taalkundige patronen te herkennen zonder dat men van te voren de exacte waarden van de entiteiten hoeft te kennen die daarin voorkomen. Men kan zo dus "zoeken zonder van te voren precies te weten wat men zoekt".

FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI) AND GETS SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON. RABBATI RECRUITED JAN ANTON KRACZEWKI (AL-KIELBASA) TO WORK FOR HIM. KRACZEWKI IS APPROXIMATELY 53 YEARS OLD, AND 180 CENTIMETERS (CM) TALL. HE DRIVES A FOUR-DOOR 1984 GREEN SUBARU. KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS.

Person	FAWIZ AL (RABBATI), RABBATI, JAN ANTON KRACZEWKI (AL-KIELBASA), KRACZEWKI
Vehicle	TEN 1-TON TRUCKS, FOUR-DOOR 1984 GREEN SUBARU
Person_Common	SMUGGLERS, ELECTRICIAN
Measure	10-15 KILOMETERS, 180 CENTIMETERS (CM)
City	KHASON
Weapon	SOPHISTICATED BOMBS
Buy Artifact	FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI)
Travel across Border	SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON
Recruit	RABBATI RECRUITED JAN ANTON KRACZEWKI (AL-KIELBASA))
Person Appearance: Age	KRACZEWKI IS APPROXIMATELY 53 YEARS OLD
Person Appearance: Height	KRACZEWKI IS 180 CENTIMETERS (CM) TALL
Person Attributes: Vehicle	HE (KRACZEWKI) DRIVES A FOUR-DOOR 1984 GREEN SUBARU
Make Artifact	KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS

Figuur 8: Voorbeeld van een analyse van entiteiten en patronen voor een typische antiterrorisme toepassing. (Bron: Inxight Software, Inc.).

Vergelijkbare voorbeelden kunnen gegeven worden voor fraudeopsporing, analyse van grote internationale en complexe criminele organisaties, en bijvoorbeeld het onderzoeken en vervolgen van oorlogsmisdaden bij de internationale gerechtshoven.

Voor dit soort toepassingen van *text mining* is brede interesse en min of meer noodzaak in de moderne maatschappij.

6.2 Sentiment mining en business intelligence

Maar er zijn ook andere gebieden waar text mining technologie relevant is. Denk aan *sentiment mining*: voor bedrijven en organisaties is het steeds vaker van belang te weten wat er positief en vooral negatief over hen geschreven wordt op het internet. Simpelweg zoeken op de eigen bedrijfsnaam is er niet meer bij: dan krijgt men teveel hits. Zoeken op alle mogelijke negatieve uitingen is ook niet te doen: er zijn gewoon teveel mogelijkheden. *Text mining, en sentiment mining* in het bijzonder bieden een uitkomst: definieer patronen van positieve en negatieve uitingen en laat *web crawlers* daarop zoeken. Dit soort technieken wordt momenteel veel gebruikt bij het vroegtijdig signaleren van potentiële PR problemen na een product introductie.

Een voorbeeld is hieronder te vinden, waarbij aan de hand van patronen van woorden bepaalt kan worden of in een recensie een positieve, negatieve of neutrale mening over een film gegeven wordt.

Figuur 9: Bron: Twitter Movie Reviews. www.twittercritic.com

Er zijn diverse *open source* woordenlijsten en semantische modellen beschikbaar die het sentiment van woorden en zinsdelen weergeven.

Deze vorm van *text mining* wordt ook veel gebruikt bij het analyseren van meningen in *blogs*, nieuwsgroepen, websites, sociale netwerken en andere internet bronnen voor bijvoorbeeld aandelen, nieuwe producten, het meten van de kwaliteit van klantenservice, en het analyseren van de mate van tevredenheid van hotelgasten.

Natuurlijk kan men naast informatie over de eigen organisatie ook informatie verzamelen over collega's en concurrenten in de markt: dit wordt ook wel *business intelligence* genoemd. *Text mining* technologie wordt hier de laatste jaren steeds vaker toegepast

6.3 Klinisch onderzoek en andere biomedische toepassingen

De noodzaak om te kunnen zoeken naar patronen waarvan men van te voren niet precies weet hoe ze eruit zien, komt ook veel voor in de farmaceutische industrie. Bij het onderzoek naar de effecten van nieuwe medicijnen of behandelingen wil men uit tienduizenden medische observaties (vaak voor een groot deel tekst documenten), patronen halen over bepaalde bijwerkingen. Ook daar is het onmogelijk om van te voren alle mogelijke voorkomens van woorden te bepalen waar men op zou willen zoeken of waar men een *alert* op zou willen zetten.

Er staan diverse voorbeelden in de literatuur waarbij problemen met nieuwe behandelingen in een vroeg stadium ontdekt konden worden met behulp van *text mining* technologie waardoor veel geld aan nutteloos onderzoek bespaard kon worden.

Andere toepassingen zijn het analyseren van medische wetenschappelijke publicaties. Dit geeft de mogelijkheid om bepaalde trends te analyseren en te voorspellen of te bepalen wie de belangrijkste auteurs, en daarmee leiders, zijn op een bepaald medisch vakgebied. Een iets minder ethische toepassing is dat sommige farmaceutische bedrijven deze "leiders" in sommige gevallen dan proberen te werven als lobbyisten voor hun medicijnen en behandelingen.

6.4 Garantieproblemen voorkomen

Eén van de eerste succesvolle commerciële toepassingen van *text mining* binnen het bedrijfsleven was het analyseren van garantieproblemen in de auto-industrie en voor consumentenelektronica. De toepassing hier bestaat uit het analyseren van reparatie rapporten van dealers zodat men vroegtijdig terugkomende patronen van garantie problemen kan ontdekken. Dit soort problemen resulteert namelijk in gratis product reparaties of soms zelfs gratis vervangingen. Des te eerder men aanpassingen in het productieproces kan maken om deze problemen te voorkomen, des te beter.

Vaak worden patronen uit interne reparatie rapporten gecombineerd met patronen van consumentenmeningen op internet en de email communicatie bij een helpdesk of een internet gebruikersgroep. Er zijn vele succesverhalen uit de praktijk waarbij met behulp van *text mining* miljoenen aan garantiekosten bespaard zijn.

6.5 Spam filters

Text mining technologie wordt ook gebruikt door spam filters die aan de hand van diverse karakteristieken van een email bericht bepalen of een bericht spam of ander ongewenst materiaal is. Omdat alleen filteren op enkele woorden vaak onvoldoende is en omdat de verzenders van spam steeds nieuwe technieken verzinnen om spam filters te omzeilen, is *text mining* technologie een krachtig nieuw gereedschap.

6.6 De kredietcrisis: e-discovery, compliance, faillissementen en data rooms

De komende jaren zal één van de grootste toepassingen van text mining gevonden worden in twee vrij nieuwe gebieden: *e-discovery* en *compliance*. Hieraan gerelateerd zijn aanverwante gebieden zoals de afhandeling van faillissementen, *due dilligence* processen en het omgaan met een *data rooms* bij een overname of fusie.

6.6.1 E-discovery

Op dit moment hebben financiële instellingen vele problemen als gevolg van de kredietcrisis. Bij twee daarvan kan *text mining* van pas komen om de kosten van onderzoeken en juridische procedures te beperken.

Ten eerste willen toezichthouders precies weten wat er verkeerd is gegaan en wie er schuldig waren. Wisten bedrijven bijvoorbeeld al in een vroeg stadium wat er aan de hand was en zijn ze willens en wetens op het verkeerde pad verder gegaan?

Het grote probleem bij het beantwoorden van vragen van toezichthouders is dat men precies moet weten wat er binnen de eigen organisatie gebeurd is en dat men vaak gevraagd wordt om, op straffe van hoge boetes of gevangenisstraffen, voor een bepaalde datum informatie te verschaffen over bepaalde soorten transacties of constructies. Omdat het lastig te bepalen is waar men dan precies op moet zoeken, zit er vaak

niets anders op dan alle beschikbare informatie door te laten lezen door specialisten. Dit is natuurlijk te duur en duurt vaak veel te lang.

Met behulp van *text mining* technologie is het makkelijker om binnen de gestelde termijnen relevante informatie aan te leveren door patronen van interesse te definiëren en de computer dit soort patronen te laten identificeren en als ze gevonden worden, hierop verder te zoeken.

Daarnaast klagen aandeelhouders en andere gedupeerden massaal financiële instellingen en andere betrokken organisaties aan. Binnen het Amerikaanse recht is het dan mogelijk om bij een tegenpartij alle potentieel relevante informatie op te vragen: een zogenaamde *subpoena* waarna een *discovery* proces volgt. Deze wetgeving is niet alleen van toepassing op Amerikaanse bedrijven, maar op *iedere* organisatie die direct of indirect zaken doet in de Verenigde Staten.

Tien tot twintig jaar geleden was er nog niet zoveel elektronische informatie als nu en in veel gevallen was het bij een *discovery* voldoende om beperkte hoeveelheden papieren informatie te onderzoeken of over te dragen.

Een extra complicatie bij een *e-discovery* vormen confidentiële gegevens: voor er een overdracht van informatie aan een derde partij plaats kan vinden, dienen eerst alle vertrouwelijke en zogenaamde *privileged* gegevens uit een collectie verwijderd of geanonimiseerd (*redaction*) te worden. Ook hier weet men van te voren vaak niet waar men op moet zoeken: sofinummers, medische dossiers van werknemers, correspondentie tussen advocaat en cliënt, vertrouwelijke technische informatie van een leverancier of klant, etc.

Men moet dus documenten zoeken waarvan men niet precies weet wat erin staat en waar ze zich precies bevinden. Vaak wordt dan teruggevallen op een lineaire *legal review* door een (duur) advocaten kantoor. De kosten hiervan lopen al snel in de miljoenen.

Door het toepassing van *text mining* kan men zeer grote besparingen realiseren. Een aanzienlijk deel van de legal review kan men dan namelijk automatisch laten plaatsvinden. Daarnaast is het met behulp van *text mining* mogelijk om snel een *early-case assessment* te maken en in te schatten hoe groot de problemen echt zijn. Dit kan belangrijk zijn als men in een vroeg stadium een schikking wil treffen.

6.6.2 Due dilligence

In deze context is de toepassing *due dilligence* (het analyseren van relevante bedrijfsgegevens bij een overname) ook van belang. Bij een *due dilligence* worden vaak *data rooms* ingericht met vele honderdduizenden pagina's met relevante contracten, financiële analyses, budgetten, etc.

In veel gevallen moet een koper in een zeer korte periode besluiten of men een bedrijf wil overnemen of niet. Vaak is het onmogelijk om alle gegevens in een *data room* binnen de gegeven tijd voldoende te analyseren. *Text mining* technologie kan hier hulp bieden.

6.6.3 Faillissementen

Een andere toepassing die meer en meer gezien wordt, is ondersteuning van een curator bij grote *faillissementen.* In veel gevallen moet een curator in een zeer kort tijdsbestek bepalen of het bestuur van een failliete onderneming alle crediteuren (inclusief henzelf) gelijk behandeld heeft (dus niet de eigen salarissen wel betalen en die van het personeel niet) en moet de curator onderzoeken of er andere onregelmatigheden zijn.

Ook bij faillissementen is steeds vaker het grootste gedeelte van alle beschikbare informatie ongestructureerde email, harde schijven vol data en andere soortgelijke gegevens.

6.6.4 Compliance, auditing en interne risico analyses

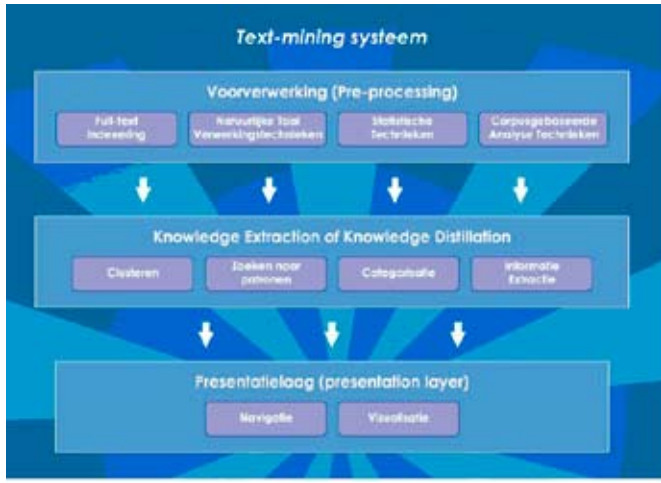
Als laatste toepassing in deze context zullen we in de toekomst zien dat door verdergaande wetgeving en strengere controlesystemen die ongetwijfeld op korte termijn zullen worden doorgevoerd, bedrijven steeds vaker (*real-time*) intern preventief onderzoek en meer diepgaande audits en risicoanalyses zullen moeten uitvoeren. *Text mining* technologie zal daarbij een onmisbaar instrument worden om op tijd in te kunnen grijpen en om de geweldige hoeveelheden informatie op tijd te kunnen verwerken en analyseren.

7 De Technologie achter Text Mining

7.1 Introductie

Een typisch text mining systeem bestaat uit de volgende onderdelen:

1. Een eerste subsysteem waarin de voorverwerking *(preprocessing)* van de gegevens plaatsvindt alvorens ze kunnen worden bewerkt, verrijkt, gevisualiseerd en geanalyseerd. Dit zijn onderdelen als *full-text* indexering, natuurlijke taal verwerking *(NLP: natural language processing)* technieken, statistische technieken en corpusgebaseerde analyse technieken.
2. Een tweede subsysteem waarin de daadwerkelijke *(core) text mining* operaties plaatsvinden als clusteren, zoeken naar patronen, categorisatie en informatie extractie. Dit wordt ook wel *knowledge* extraction of *knowledge distillation* genoemd.
3. Een derde subsysteem bestaat uit de presentatielaag *(presentation layer)* ten behoeve van de gebruikers van het systeem met onder andere navigatie, visualisatie en andere technieken om gegevens daadwerkelijk te analyseren en eventueel handmatig verder te verrijken en organiseren, dan wel om een kwaliteitscontrole uit te voeren.



Figuur 10: Een typisch text mining systeem

Hieronder zullen deze drie subsystemen één voor één in detail besproken worden.

7.2 Preprocessing

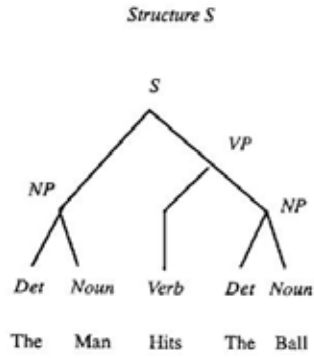
In de eerste fase van *text mining* zullen een aantal basis stappen ondernomen moeten worden. Het doel van deze voorbewerkingen is om de volledig ongestructureerde en vaak pluriforme informatie terug te brengen tot een gemeenschappelijke noemer. Een aantal vormen van voorbewerking zijn (volgorde is willekeurig):

- Scannen van papier naar digitale bestanden.
- Herkennen van tekst van een bitmap representatie *(optical character recognition: OCR)*.
- Uitpakken van gecomprimeerde en samengestelde bestanden (ZIP, Email).
- Herkennen van het formaat en karakterset van een elektronisch bestand (PDF, HTML, ASCII, ANSI, UNICODE, MS-Word, etc.).
- Herkennen van de spraakcomponent van een multimediaal bestand.
- Extraheren van tekst (in de goede volgorde) uit een document.
- Herkennen van de taal van een document, pagina of paragraaf.
- Machinaal vertalen van de inhoud van een bestand.
- Automatisch maken van samenvattingen.
- Verwijderen van woorden *(tokens)*, ruiswoorden, punctuering, en andere leestekens.
- Bouwen van een *inverted-file full-text* index.
- Herkennen van taalkundige zinnen.
- Herkennen van woordstammen en verwijderen van vervoegingen.
- Bepalen van de grammaticale structuur van een zin. Dit kan zowel met statistische als met taalkundige technieken.
- Herkennen van taalkundige verwijzigen *("de man loopt met zijn hond, hij ziet een vliegtuig vliegen")*.
- Herkennen van zogenaamde eigennamen *(named entities)*.
- Herkennen van synoniemen, homoniemen, afkortingen, uitdrukkingen (idioom) en andere taalkundige varianten.

Vaak kunnen per subtaak verschillende technieken gebruikt worden. Zo is het mogelijk om de grammaticale analyse met zowel grammaticale technieken (klassieke zinontleding met regels), als met statistische technieken (*hidden-Markov* ketens en andere technieken uit de *machine learning*), als met corpusgebaseerde technieken, of zelfs een combinatie van bovenstaande technieken uit te voeren.

Voor *text mining* is in veel gevallen echter niet een zeer diepgaande analyse nodig, zodat volstaan kan worden met een redelijke oppervlakkige analyse waarbij de belangrijkste elementen van een zin worden herkend: de onderwerpszin, de werkwoordzin, potentiële eigennamen, verwijzingen, en andere relaties. In veel gevallen worden *finit- state parsers* of *shallow* parsers gebruikt met ondersteuning van woordenboeken. Deze analyse wordt ook wel vaak een *part-of-speech* (POS) analyse genoemd.

Met behulp van een corpus (een grote collectie van voorgeanalyseerde data) is het vervolgens mogelijk om statistische waarschijnlijkheden van bepaalde taalkundige voorkomens mee te geven.



Figuur 11: Een part-of-speech analyse van een eenvoudige zin (Bron: Scholtes, 1993)

Het resultaat van de *preprocessing* fase is een uniform gegevensformaat waarbij de tekst voorzien is van diverse statistische-, regelgebaseerde- en taalkundige kenmerken die nodig zijn voor de volgende fase: de *core text mining*.

In deze fase moet altijd een balans gevonden worden tussen de kwaliteit en kwantiteit van de aanvullende verrijkingen aan de ene kant en snelheid waarmee deze plaatsvindt aan de andere kant. Vaak moeten namelijk gigabytes of zelfs terabytes aan tekst verwerkt worden en dat moet wel binnen acceptabele rekentijden plaatsvinden.

Belangrijk is ook dat de technieken die gebruikt worden robuust zijn (om kunnen gaan met gegevens waar fouten en onbekendheden in zitten: spelfouten, scanfouten, schrijffouten, typefouten, nieuw jargon, onbekende afkortingen, etc.). Vaak prevaleren statistische- en corpus gebaseerde technieken daarom boven regelgebaseerde grammaticale technieken, mede omdat de eerste twee robuuster en sneller zijn. De beschikbaarheid van grote (*open source*) corpora en gigantische hoeveelheden digitale teksten op het internet maakt de toepassing van deze technieken de laatste tijd alleen maar logischer en makkelijker.

7.3 Core text mining

In de *core text mining* fase vindt de zogenaamde *knowledge discovery of knowledge distillation* plaats. Hierbij worden entiteiten (voornamelijk eigen woorden) geclassificeerd, patronen of sentimenten worden herkend, en documenten kunnen worden geclusterd of gecategoriseerd. Hieronder zal kort worden ingegaan op deze verschillende technieken voor zover die nog niet eerder besproken zijn.

Veel van deze technieken komen uit de klassieke patroonherkenning, alleen bestaat de invoer bij *text mining* uit tekstuele data die eerst moet worden omgezet in getallen. Dit omzetten van andersoortige data naar getallen wordt ook wel eigenschap selectie *(feature selection)* en eigenschap extractie *(feature extraction)* genoemd. De kunst is vaak de beste eigenschappen van een bepaald object te selecteren en die dan te meten en vervolgens de meest onderscheidende eigenschappen te extraheren (vaak is dit een dimensie reductie).

Dit is een probleem dat niet alleen in de *text mining* speelt, maar ook bij spraakherkenning, beeldverwerking en andere toepassingen waarbij een computer niet om kan gaan met de oorspronkelijke analoge data: met behulp van slimme technieken moeten de gegevens eerst vertaald worden naar een wiskundige representatie. Vervolgens kunnen dan de traditionele patroonherkenning algoritmes gebruikt worden zoals *probabilistic classifiers, tree classifiers, decision rule classifier,* neurale netwerken, clustering *(k-means, nearest neighbour, gather/scather), hidden Markov models*, etc.

In het geval van *text mining* worden deze patroonherkenningstechnieken dus toegepast op de oorspronkelijke tekstuele informatie, op taalkundige

eigenschappen die in de preprocessing fase zijn afgeleid en op andere contextuele informatie die van belang kan zijn. Dit kan dus informatie zijn die zowel impliciet als expliciet aanwezig is. Het laatste wordt ook wel eens domeinkennis genoemd: aanvullende kennis over een specifiek probleem, vakgebied of (tijdelijke) situatie.

7.3.1 Informatie extractie

Bij het extraheren van informatie kunnen de volgende basis elementen in een tekst herkend worden:

1. *Entiteiten:* de basiseenheden die in een tekst gevonden kunnen worden. Bijvoorbeeld: mensen, bedrijven, locaties, producten, medicijnen, en genen.
2. *Attributen:* dit zijn eigenschappen van de gevonden entiteiten: denk aan functienamen, leeftijden en sofinummers van personen, adressen van locaties, bedragen van producten, kentekens van auto's en het type organisatie.
3. *Feiten:* dit zijn relaties tussen entiteiten. Bijvoorbeeld een arbeidsrelatie tussen een bedrijf en een persoon.
4. *Gebeurtenissen:* dit zijn interessante gebeurtenissen of activiteiten waarin entiteiten zijn betrokken zoals: "een persoon praat met een ander persoon", "een persoon reist naar een locatie", en "een bedrijf maakt geld over aan een ander bedrijf".

7.3.1.1 Entiteiten en attributen

Het eerste onderzoek naar de zogenaamde *named entity extraction* stamt al uit een door het Amerikaanse *Defense Advanced Research Project Agency* (DARPA) gesubsidieerd onderzoek dat in 1995 werd uitgevoerd onder de vlag van de *Message Understanding Conference (MUC-6).* Eén van de taken van dit onderzoek was om in vrije tekst (vaak berichtenverkeer of openbare nieuwsberichten) alle voorkomende personen, locaties, organisaties, tijden en aantallen te herkennen. Omdat men van te voren niet wist wat voor eigen kenmerken er in de tekst zouden voorkomen was het noodzakelijk om eerst een taalkundige analyse van de tekst te maken, en vervolgens kon men daarmee de eigen kenmerken *(named entities)* identificeren en deze daarna aan de hand van verschillende technieken classificeren in mogelijke categorieën.

Eén van de manieren om dit te doen is met behulp van reguliere expressies. Hiermee kunnen data, telefoonnummers, internetadressen, bankrekeningnummers en sofinummers redelijk goed herkend worden.

Een goed voorbeeld van een reguliere expressie om een email adres te vinden is:

```
\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
```

Figuur 12: Voorbeeld van een reguliere expressie (Bron: http://www.regular-expressions.info/examples.html)

Het is vrij complex en ook veel werk om dit soort reguliere expressies te definiëren, vooral omdat er veel varianten van patronen kunnen voorkomen en men niet altijd alles met één eenvoudige reguliere expressie kan omvatten: men krijgt òf hele complexe patronen of er zijn altijd wel entiteiten die door de vele onregelmatigheden niet met reguliere expressies te herkennen zijn. Er moeten (de naam zegt het al) veel gelijksoortige (reguliere) patronen in entiteiten zitten om ze met deze technologie goed te kunnen classificeren. Er is echter een recent boek dat de titel draagt: *Practical Text mining with Perl,* waarin men het gebruik van reguliere expressies tot het extreme doorvoert, en vervolgens een heel eind komt [Bilisoly, 2008].

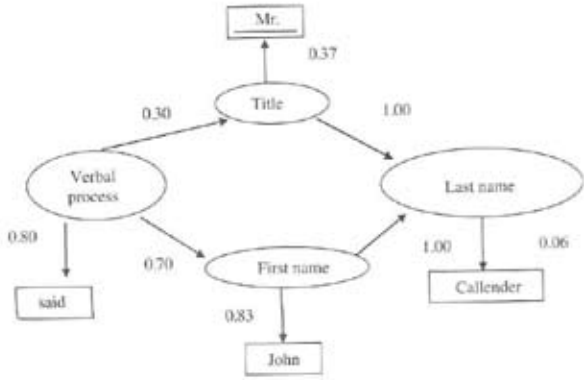
Een logische andere aanpak is om de langst voorkomende vorm van een *named entity* te vergelijken met bekende *named entities* in woordenboeken, waarbij in het woordenboek vervolgens wordt bijgehouden wat voor soort entiteit een bepaald woord of een bepaalde combinatie van woorden is.

Hierbij kan ook rekening gehouden worden met de taalkundige waarschijnlijkheid dat een bepaald woord een bepaalde betekenis heeft. Dit is waar *hidden-Markov* modellen *(HMM)* een belangrijke rol spelen.

Binnen de *text mining* wordt met een *hidden-Markov* model weergegeven wat de kans is dat na bijvoorbeeld een aanspreektitel zoals Mr., Meneer, Mevrouw of Dr. een achternaam komt. Diverse relaties tussen woorden en hun context kunnen op deze manier formeel worden vastgelegd. De waarschijnlijkheden zijn automatisch af te leiden uit grote corpora met voorbeeldteksten. Aan de hand van een dergelijk model is het mogelijk om de taalkundige waarschijnlijkheid te bepalen of een entiteit bijvoorbeeld

een locatie is of een persoonsnaam. Een goed voorbeeld in deze context is de entiteit "Mr. Holland", waarbij het voor mensen direct duidelijk is dat het hier geen locatie maar een persoon betreft. Met behulp van een *hidden-Markov* model kan een algoritme ook snel dezelfde goede beslissing nemen.

Het grote voordeel van deze techniek is dat de kennis van een onderliggende taal minimaal hoeft te zijn.

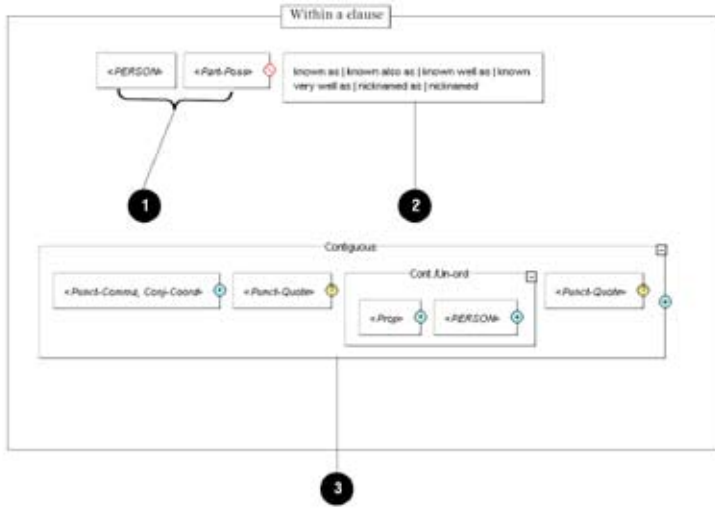


Figuur 13: Een voorbeeld van een hidden-Markov model voor named-entity recognition (Bron: Moens, 2006).

Vanzelfsprekend is het ook mogelijk om de bovenstaande technieken te combineren en vervolgens de meest waarschijnlijke classificatie aan een entiteit toe te kennen. Vaak wordt meer dan 90% van de aanwezige attributen en entiteiten met een combinatie van de hier vermelde technieken herkend.

7.3.1.2 Feiten

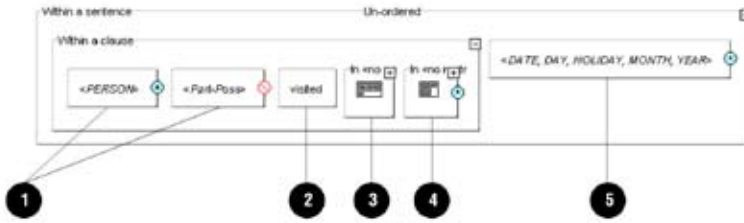
Bij het herkennen van feiten (relaties tussen entiteiten en hun attributen) kunnen regels nuttig zijn. Zo is hieronder een model weergegeven dat, aan de hand van al herkende entiteiten als persoonsnamen, in staat is om met behulp van de taalkundige context synoniemen of aliassen te herkennen.



Figuur 14: Een regel om voor persoonsnamen aliassen te herkennen (Bron: Inxight Software Inc.).

7.3.1.3 Gebeurtenissen

Bij het ontdekken van een gebeurtenis worden relaties tussen entiteiten herkend. Dit is één van de meest interessante vormen van patroonherkenning omdat het zeer complexe patronen kan herkennen zoals: "een persoon praat met een ander persoon", "een persoon reist naar een locatie", en "een bedrijf maakt geld over aan een ander bedrijf".



Figuur 15: Een regel om te ontdekken wie wie bezoekt op welke dag (Bron: Inxight Software Inc.).

Een van de grootste problemen bij het ontdekken en herkennen van gebeurtenissen is het oplossen van zogenaamde *anaphora* en *coreferences*. Dit is het taalkundige probleem om paren van taalkundige uitdrukkingen te kunnen linken die verwijzen naar dezelfde entiteiten

uit de echte wereld. In MUC-6 (1995) and MUC-7 (1998) is er voor het eerst onderzoek gedaan naar deze problemen.

Denk bijvoorbeeld aan de volgende tekst:

> *"Een man loopt naar het station en probeert de trein te halen. Zijn naam is Jan Jansen. Even later ontmoet hij zijn collega, die net een kaartje voor dezelfde trein heeft gekocht. Samen zijn ze werkzaam bij de NS als technisch medewerker en ze gaan naar een bespreking met collega's in Utrecht".*

In deze tekst staan diverse verwijzingen en coreferenties. Er zijn diverse soorten van *anaphora* en co-referenties die gedisambigueerd moeten worden wil het mogelijk zijn om complexere patronen van gebeurtenissen volledig te doorgronden en uit de tekst te extraheren. Een aantal voorbeelden van dit soort (onderlinge) verwijzingen zijn:

- *Pronominal Anaphora:* hij, zij, wij, zichzelf, etc.
- *Proper Name Coreference:* bijvoorbeeld meerdere referenties naar dezelfde naam.
- *Apposition:* het geven van aanvullende informatie op een entiteit, zoals "Jan Jansen, de vader van Piet Jansen".
- *Predicate Nominative:* hierbij wordt een aanvullende beschrijving gegeven van een entiteit. Bijvoorbeeld: Jan Jansen, die de voorzitter is van de voetbalclub.
- *Identical Sets:* Meerdere sets van verwijzingen naar entiteiten die gelijk zijn zoals: "Ajax", "het beste team", en de "groep van spelers" refereren allemaal naar dezelfde groep personen.

Er zijn verschillende manieren om deze problemen te benaderen: (i) met een diepgaande taalkundige analyse van een zin, of (ii) aan de hand van een groot geannoteerd corpus. Beide technieken hebben hun voor en nadelen. Op dit gebied is de komende jaren nog veel onderzoek noodzakelijk om een betere kwaliteit van dit soort analyses te krijgen.

In deze context kan ook het maken van analyses in de tijd genoemd worden en het volgen van onderwerpen over meerdere documenten en door grotere collecties. Vooral bij de analyse van email collecties kan dit heel interessant zijn.

7.3.1.4 Sentimenten

Al eerder is het begrip *sentiment mining* toegelicht. Hierbij wordt aan de hand van gebruikte bijvoeglijke naam- en werkwoorden bepaald of het sentiment van een document positief, negatief of neutraal is. Dit gaat meestal aan de hand van het vergelijken van woorden die in een tekst gebruikt worden met een tabel waarin de sentiment waarden van die woorden staan.

Helaas is deze techniek niet zo betrouwbaar en ook nog niet zo vergevorderd als de hierboven beschreven extractie technieken. De komende jaren is er dan ook voldoende ruimte om ook de kwaliteit van *sentiment mining* technieken op het niveau van de entiteit extractie te krijgen.

7.3.2 Categorisatie en classificatie

Hierboven is uitgebreid ingegaan op het categoriseren en classificeren van zogenaamde *named entities*, maar men kan dit principe ook doortrekken naar het categoriseren en classificeren van gehele documenten of delen van documenten. In deze context is het nuttig om clustering van documenten te noemen.

In het algemeen kan men categorisatie-, classificatie- en clustering algoritmes verdelen in twee hoofdgroepen: *supervised* en *non-supervised* (ook wel zelforganiserend genoemd).

7.3.2.1 Supervised technieken

Supervised technieken worden van te voren getraind met een representatieve training set en kunnen daarna voor andere data gebruikt worden. Mogelijke categorieën moeten van te voren bekend zijn en worden expliciet aan het systeem geleerd in combinatie met bijbehorende invoergegevens. Het eerder genoemde *hidden-Markov* model is hier een goed voorbeeld van. Andere voorbeelden zijn supervised neurale netwerken (*back-propagation neural networks)*, stochastische contextvrije grammatica's, maximale entropie modellen en Support Vector Machines.

In alle gevallen dient men in eerste instantie relevante eigenschappen van documenten af te leiden om deze vervolgens te gebruiken om de bovengenoemde algoritmes te trainen.

7.3.2.2 Un-supervised technieken

Bij *un-supervised* of zelforganiserende technieken wordt een grote hoeveelheid representatieve data aan het systeem gepresenteerd, waarna het betreffende algoritme of model zelf de data herkent, analyseert, organiseert en ervan leert, zodat nieuwe data in de toekomst aan de hand van hetzelfde model automatisch geclassificeerd kan worden. Categorieën zijn van te voren niet bekend; het systeem herkent ze zelf. Het voordeel van zelforganiserende modellen is dat er geen training vereist is. Het nadeel is dat convergentie naar een stabiele, correcte of zelf optimale toestand niet altijd gegarandeerd is.

Bij al deze technieken wordt eerst een bepaalde wiskundige afstand gedefinieerd (Euclidisch, Cosinus, Levenstein, City Block, etc.) die vervolgens wordt toegepast op een set van vectoren met getallen, die op hun beurt weer eigenschappen van documenten representeren. In sommige gevallen zijn dit (stam)woorden, in andere gevallen zijn het semantische groepen *(Latent Semantic Indexing: LSI)* of het zijn bijvoorbeeld de eerder beschreven herkende entiteiten, attributen, feiten of gebeurtenissen. LSI is een goede techniek om de dimensie van de vectoren te reduceren, net als *principle commonent* analyse en andere vergelijkbare dimensie reductie technieken uit de wiskunde.

De vectoren (en daarmee indirect de documenten of begrippen) worden dan conform de gekozen wiskundige afstand ten opzichte van elkaar georganiseerd of geclassificeerd. Dit is het mogelijk door de vectoren met cluster- en zelforganiserende algoritmes te clusteren en hieruit bijvoorbeeld automatisch relaties tussen begrippen, entiteiten, of concepten af te leiden.

Clustering is ook zeer nuttig voor het herkennen van groepen van duplicaten, hoewel het in veel gevallen niet echt praktisch is omdat de computationele complexiteit van cluster algoritmes in het algemeen kwadratisch is met het aantal documenten in de te clusteren set. Daardoor is clustering qua complexiteit bij een set van bijvoorbeeld 10 miljoen email bestanden niet echt meer toepasbaar.

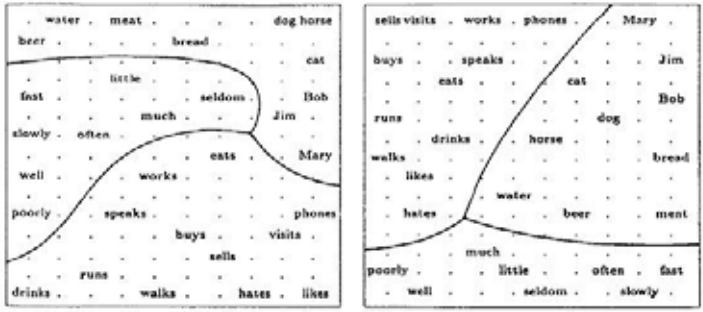


FIGURE 11. The semantotopic map for input manually tagged with contextual information (left) and the map for a restricted context of the immediate predecessor only (right) (reprinted from [Ritter et al., 1989b]).

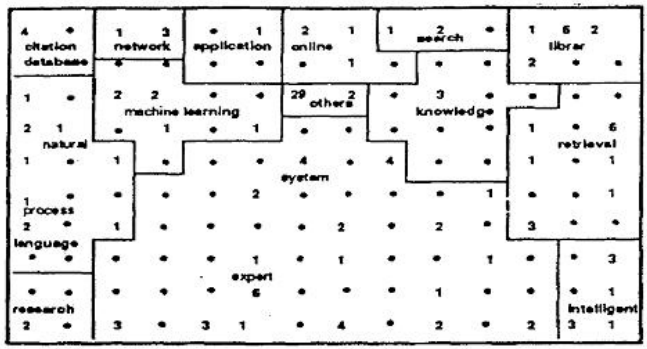
Figuur 16: Een voorbeeld van het clusteren van woorden in semantische groepen met een zelforganiserend neuraal netwerk (Bron: Scholtes, 1993).

Vergelijkbare technieken kunnen gebruikt worden om automatisch een taxonomie af te leiden uit ongestructureerde document collecties.

Voorbeelden van zelforganiserende technieken zijn *Kohonen* zelforganiserende neurale netwerken en diverse andere cluster technieken zoals *N-Nearest Neighbour, K-Means* en het binnen de *text mining* populaire *Scatter/Gather* algoritme.

Bij het *Scatter/Gather* algoritme wordt optimaal gebruik gemaakt van de combinatie van handmatig bladeren en machinaal clusteren. In eerste instantie worden documenten gevonden door middel van woorden in een *full-text* index. Echter, indien meer algemene vragen gesteld worden, zal teruggevallen worden op een logische inhoudsopgave en zullen "naburige" documenten gepresenteerd worden.

Bij iedere iteratie in een *Scatter/Gather* sessie, wordt een document collectie in eerste instantie verdeeld *(scatter)* in sets van clusters en de korte beschrijving van de clusters wordt aan de gebruikers gepresenteerd. Van deze beschrijving wordt dan een nieuwe subcollectie gemaakt *(gather)*. Hierop wordt het *Scatter/Gather* process dan nog een keer herhaald, net zolang tot er voldoende resolutie is. Op deze manier zal een dynamische inhoudsopgave gemaakt worden die gebruikt kan worden bij het navigeren door de documenten.



A self-organizing semantic map of AI literature. 140 documents from LISA database are used as input to produce the map. The areas on the map are automatically generated, their relative positions, neighbors, and sizes are determined by the input data. The numbers on the map represent the number of documents mapped to each node.

Figuur 17: Voorbeeld van een zelforganiserend neuraal netwerk waarbij Artificial Intelligence publicaties automatisch georganiseerd zijn op onderwerp (Bron Scholtes, 1994b).

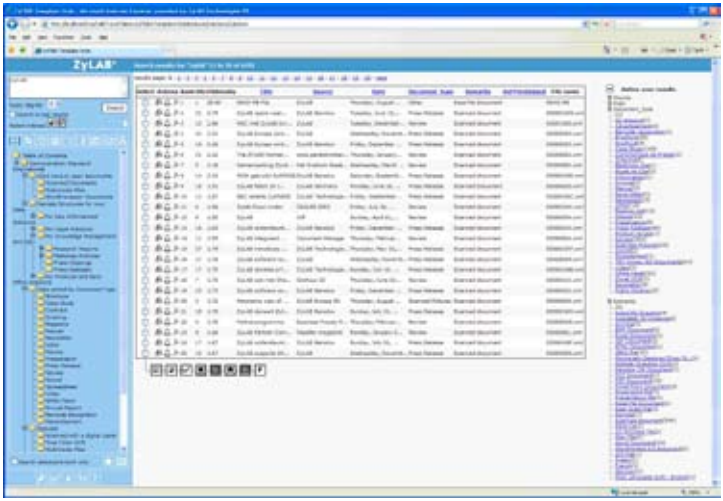
In de praktijk zijn clustering van documenten en het automatisch afleiden van een taxonomie niet erg succesvol. Dit komt voornamelijk omdat de kwaliteit vaak maar voor 50% correct is en de rest handmatige aanpassingen vereist. Ook zijn er veel voorbeelden van systemen die de ene keer wel goed convergeren en de andere keer niet of die iedere iteratie andere uitkomsten geven. *Un-supervised* classificatie, clustering en automatische taxonomie generatie systemen vereisen in alle gevallen minimaal enige menselijke interventie zoals bij het *Scatter/Gather* algoritme en dan kunnen vaak wel redelijke successen behaald worden.

7.4 Presentatie laag van een text mining systeem

Na alle bewerkingen zoals die zijn gepresenteerd in de vorige secties, is de oorspronkelijke data voorzien van diverse aanvullende eigenschappen. Hierdoor komt een volledig nieuw scala aan analyse en zoektechnieken ter beschikking.

Hiervan kan gebruik gemaakt worden in de presentatie laag van het *text mining* systeem.

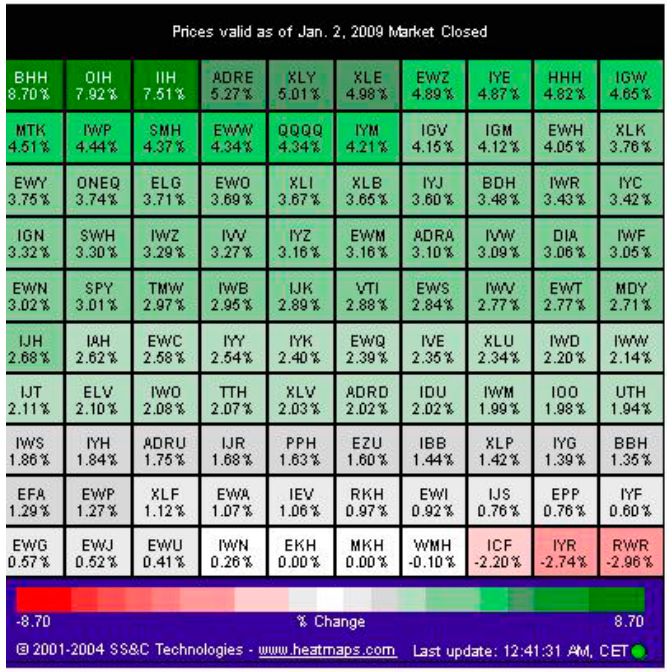
Zo is het met de verrijkte data mogelijk data te visualiseren (zie eerder), complexe statistische analyses te maken, vergelijkbare documenten op te roepen tijdens het zoeken, op kenmerken door te zoeken, op kenmerken te clusteren, te navigeren aan de hand van de volledige tekst van een document én aan de hand van de vele beschikbare kenmerken van een document, etc.



Figuur 18: Presentatie en de mogelijkheid tot organiseren, navigeren en doorzoeken op (automatisch gevonden) kenmerken van eerder gevonden documenten (bron: ZyLAB Technologies BV).

Binnen deze context is het vanzelfsprekend heel belangrijk om een goede, intuïtieve en duidelijke gebruikersinterface te hebben om van alle nieuwe zoekmogelijkheden gebruik te maken.

Een bijzondere manier van visualisatie is het nalopen van links tussen documenten (email berichten in het bijzonder) of het maken van tijdslijnen waarop documenten gerepresenteerd worden of het genereren van zogenaamde *heat maps* om veranderingen of onderlinge relaties tussen documenten weer te geven.



Figuur 19: Heatmap van de NASDAQ Stock Exchange op 2 januari 2009 (bron: www.nasdaq.com).

8 Onderwijs en Onderzoek

Een kort woord over het onderwijs en onderzoek: onderwijs en onderzoek zijn de kerntaken van de leerstoel *text mining*.

De leerstoel zal zich hierbij richten op het onderwijzen van *text mining* methodieken voor taalafhankelijke *feature selection* en *feature extraction* zodat documenten kunnen worden voorzien van diverse extra entiteiten, attributen, feiten, gebeurtenissen, sentimenten en relaties die met behulp van geavanceerde gebruikersinterfaces goed doorzocht, gevisualiseerd, geanalyseerd en gefilterd kunnen worden.

In februari 2009, volgende maand, beginnen we met een college voor de *Masters Course on Knowledge Engineering* opleiding, genaamd *text mining*. Op termijn ligt het in de bedoeling om ook onderwijstaken te verzorgen

voor *text mining* gerelateerde (gast)colleges bij de andere faculteiten van de Universiteit Maastricht. Zoals ik eerder heb aangegeven zijn er diverse aanknopingspunten met *life sciences, governance,* forensische en vanzelfsprekend juridische toepassingen.

De colleges zullen gericht zijn op het begrijpen van de hier eerder besproken technieken en de praktische toepasbaarheid daarvan binnen diverse vakgebieden.

Hiervoor zal onder andere gebruik gemaakt worden van diverse *open-source text mining* bibliotheken waarmee studenten zelf alledaagse problemen kunnen oplossen door de toepassing van *text mining*.

In de nabije toekomst zullen samen met *Masters* en eventueel ook *Ph.D.* studenten relevante onderzoeksonderwerpen gedefinieerd worden. Zoals eerder aangegeven is het vakgebied van *text mining* een jong vakgebied met vele deelgebieden waar onderzoek mogelijk en ook gewenst is. In samenspraak met de Universiteit en gekwalificeerde derde partijen zal de komende jaren gewerkt worden aan het verder brengen van het vakgebied door het uitvoeren van relevant onderzoek.

Ook zijn er diverse internationale *text mining* onderzoeksactiviteiten zoals de Legal TREC van de *University of Maryland* en diverse initiatieven binnen de Europese Unie waarbij aansluiting gezocht zal worden.

Onderzoek naar nieuwe technieken, evenals het uitbreiden van bestaande technieken voor andere of meer talen of het maken van applicatie-templates voor snellere toepasbaarheid zijn mogelijke onderzoeksonderwerpen.

9 Conclusies en Vooruitblik

9.1 Van lezen naar zoeken en vinden

In Delft maakte ik in 1982 kennis met de beginselen van de informatica. Ik verdiepte me in statistiek, patroonherkenning, artificiële intelligentie en leerde in 1987 samen met anderen een computer programma 'lezen' via zogenaamde *optical character recognition (OCR)* technologie. Met de parallelle verwerkingskracht van de NCUBE computer die we tot onze beschikking hadden (vergelijkbaar met 16 Digital VAX computers uit

die tijd), was het op dat moment mogelijk om één pagina per minuut te verwerken. Nu, 22 jaar later, krijgt men bij een scanner van vijftig euro een gratis OCR pakket dat op een standaard PC per seconde bijna foutloos een volle pagina tekst leest.

Eind jaren tachtig, tijdens mijn militaire diensttijd bij de inlichtingendienst van de Koninklijke Marine was een paar gigabyte tekst per dag veel. Om daar zinnige dingen mee te doen, had men voor die tijd gigantische computers en opslagcapaciteit nodig. Nu lachen we daarom.

Tussen 1989 en 1993 heb ik getracht om met behulp van op neurale netwerken gebaseerde algoritmes, computers de structuur van taal te leren, structuren van taal af te leiden en in grote collecties tekst te zoeken. De toepassingen waren divers: van machinaal vertalen tot spraakherkenning en zoekmachines. Maar commercieel toepasbaar waren de technieken op dat moment nog niet.

Wel werden op dat moment de algoritmes die Gerald Salton eind van de jaren zestig ontwikkeld had voor het zoeken in grote hoeveelheden tekst net beschikbaar op het PC-platform. Dit was het begin van een commerciële revolutie en ook het begin van het succes van ZyLAB.

In de jaren negentig begon ook het onderzoek naar *text mining*. Op universiteiten en in samenwerking met DARPA (*TREC en de Message Understanding Conferences)* werden slimme algoritmes gemaakt om teksten samen te vatten, entiteiten te extraheren, documenten te clusteren, data te visualiseren, etc. Er werd veel onderzoek gedaan naar machinaal vertalen en spraakherkenning. Maar ook hier was de commerciële toepasbaarheid toen nog ver te zoeken.

Sinds 2000 is *full-text* zoeken een groot gemeengoed geworden. Iedereen gebruikt internet zoekmachines en de onderliggende technologie is algemeen geaccepteerd. Hoe anders was dat begin jaren negentig, toen het aan de man (of vrouw) brengen van *full-text retrieval* technologie nog gelijk stond aan evangeliseren. Menig bibliothecaris rilde van het idee om een eindgebruiker *full-text* toegang te geven tot een collectie: dat kon niet en was veel te gevaarlijk!

Tegenwoordig kunnen we niet meer zonder: we hebben toegang tot de inhoud van volledige bibliotheek en min of meer tot bijna alles wat ooit

in de wereld geschreven is via het internet. We kunnen razendsnel zoeken en (denken alles te) vinden wat we zoeken. Nu zien we dat de balans zelfs is doorgeslagen: veel mensen denken dat we met de beschikbare internet zoekmachines alle zoekproblemen hebben opgelost. Ik hoop dat ik vandaag duidelijk heb kunnen maken dat dat niet altijd zo is.

We kunnen in ieder geval vaststellen dat op dit moment het vakgebied van de *text mining* en machinaal vertalen duidelijk bezig zijn met een commerciële doorbraak. Successen uit de inlichtingenwereld vinden nieuwe toepassingen in juridische-, medische- en industriële toepassingen. Zelfs marktonderzoek gaat tegenwoordig voor een groot deel door het internet af te zoeken naar consumenten meningen.

9.2 De generatiekloof

Een interessante observatie in deze context is dat nieuwe technologie vaak ongeveer twintig jaar nodig heeft om tot volle wasdom te komen, we hebben dat kunnen waarnemen bij de PC (1982-2002), het internet (1990-2010), en GSM telefoons. Men kan deze lijn zelfs terugtrekken tot de uitvindingen van de telegraaf, telefoon, de radio en de TV. Dit is ook heel logisch, want die twintig jaar dat is precies een generatie. Wij mensen hebben waarschijnlijk een generatie nodig om aan nieuwe technologie te wennen en ook om het "aan te passen" aan wat we acceptabel vinden.

9.3 Gevolgen van nieuwe informatie technologie

Als gevolg van al deze nieuwe informatie technologie zijn een aantal zaken de laatste tijd radicaal veranderd:

- Neem bijvoorbeeld marketing: men stuurt tegenwoordig niet meer willekeurig folders rond, maar men wacht tot mensen op bepaalde woorden zoeken die men interessant vindt en dan koopt men daar advertentieruimte omheen.
- Opsporingsdiensten kunnen nu ook kijken wie waar op gezocht heeft en, als dat mag, preventief actie ondernemen.
- Bedrijven en organisaties laten vrijwilligers problemen oplossen waar hun eigen onderzoeksafdelingen niet uitkomen of waar geen geld voor is.

De meeste van deze hedendaagse successen zijn een gevolg van de technologie die twintig jaar geleden voor het eerst ontwikkeld werd. We worden met zijn allen efficiënter, dat staat vast. Dat moet ook, want het is onmogelijk om zonder dit soort nieuwe technieken en principes de hedendaagse informatiestromen te verwerken of te controleren.

9.4 Andere te verwachten ontwikkelingen

De komende jaren staat ons nog veel meer te wachten:

- Computers zullen meer en meer automatisch data structureren en organiseren aan de hand van de inhoud van die data. Die structuren en eigenschappen zullen gebruikt worden om ons informatie op maat aan te leveren. Door vooruitgang op gebieden als de *text mining*, ons begrip van menselijke taal, het automatisch vertalen, spraakherkenning, beeldherkenning, e.d. zal dit alleen nog maar sneller gaan.
- Informatie die aanwezig is op sociale netwerken zal gekoppeld en geïntegreerd worden. Men wordt professioneel gekoppeld aan mensen met vergelijkbare interesses. Dat wordt in sommige gevallen zelfs al met routeplanners en mobiele telefoons geïntegreerd: men krijgt een SMS als er iemand binnen 10 meter is met vergelijkbare interesses!
- Interactieve websites zullen ons meer en meer gefilterde informatie presenteren waarbij rekening gehouden wordt met wie we zijn, wat onze interesses zijn en in welke informatie mensen met vergelijkbare interesses ook geïnteresseerd waren.
- Advertenties en andere commerciële aanbiedingen zullen gerichter aangeboden worden. Computers zullen steeds meer rekening gaan houden met: locatie, cultuur, ras, rijkdom, leeftijd, seksuele voorkeur, etc. Er is veel meer informatie over ons beschikbaar dan we denken.
- Taalbarrières zullen verdwijnen. De kwaliteit en binnenkort ook de gratis beschikbaarheid van hoge kwaliteit machinale vertaalsystemen zal het mogelijk maken om dynamisch informatie uit de hele wereld te vertalen en te gebruiken.
- De vorm van informatie wordt volledig transparant: telefoon, geluid, video, plaatjes, tekst, alles wordt geïntegreerd en alles wordt doorzoekbaar.

Een van de redenen waarom ik denk dat ons dit allemaal op redelijk korte termijn te wachten staat is de ontwikkeling van *The Grid:* dit is een groot gedistribueerd (vaak wereldwijd) netwerk van heel veel kleine computers. Denk aan honderdduizend tot meer dan een miljoen computers. Google en Microsoft bouwen al jaren aan een dergelijk eigen netwerk: miljarden geven ze er aan uit. Daarop is alles aan elkaar gekoppeld (ook al onze informatie). Een bekende formule is dat een netwerk even krachtig is als het kwadraat van het aantal gebruikers of computers in dat netwerk. Er staat ons dus nog heel wat te wachten!

The Grid is de infrastructuur achter een soort "wolk" (*the Cloud* wordt het ook wel in het Engels genoemd) die softwareprogramma's host die op meerdere computers tegelijk en gedistribueerd kunnen draaien. Alle informatie is ook op meerdere computers tegelijk opgeslagen. Het is dus overal en nergens. Dit is een zeer krachtig concept. Er is in theorie geen eigen opslag meer nodig. Men kan overal bij met iedere computer die aan het internet gekoppeld is, maar ook met bijvoorbeeld een mobiele telefoon. Tien jaar gelden begon het onderzoek naar software die nodig was om gebruik te maken van dit soort hardware. Vele manjaren zijn besteed aan het praktisch toepasbaar maken van een *Cloud* op een *Grid* architectuur. Het lijkt erop dat dit binnenkort op grote schaal gaat lukken.

En als we daarna verder kijken (of dromen), dan is er heel ver aan de horizon nog meer. Het meest revolutionair wordt waarschijnlijk de doorbraak van de kwantum computer (als die er ooit komt). Een kwantum computer is voor bepaalde toepassingen exponentieel zo krachtig vergeleken met gewone computers. De eerste kwantum algoritmes om massaal informatie te verwerken en te doorzoeken zijn er al: zoeken in grote hoeveelheden informatie kan daarmee nog beter en vooral sneller.

Kwantum computers zijn vooral goed in het zoeken in complexe hoogdimensionale data, die ook vaak nog onvolledig is en vol ruis zit, zoals geluid, spraak, taal, video, beeld, en DNA. Het probleem is alleen dat we nog geen echte kwantum "chips" kunnen maken, we komen nog niet verder dan simulaties. Maar bij de TU-Delft doen ze revolutionair onderzoek naar echte kwantum computeronderdelen. Wellicht dat dit nog 50-100 jaar duurt, maar dat het eraan komt is zeer waarschijnlijk.

Verder onderzoek naar het omzetten van algoritmes en principes zoals we die nu kennen op binaire computers naar algoritmes die geschikt zijn

voor kwantum computers is één van de meer interessante uitdagingen voor de informatica wetenschap.

9.5 De komende twintig jaar

Zoals ik eerder aangaf, vonden we eind van de jaren tachtig twee gigabyte per dag veel informatie. Nu begint een paar terabyte (1.024 gigabyte) per dag een probleem te worden, toch is dat ook best te overzien. Maar als Moore's wetten van kracht blijven, dan hebben we over twintig jaar picabytes (1.024 terabyte) per dag te verwerken. En dat vinden we nu wel een probleem, want om daarmee om te gaan hebben we nog niet echt de technieken in huis. Ook zal de aard van de informatie de komende jaren veranderen: al die data zal, naast het bevatten van oneindig veel duplicaten, ook vooral transactioneel, real-time en multi-mediaal van aard zijn.

We hebben nu vaak al problemen met de complexe structuur van email bestanden, wat dus te denken van al die gigantische hoeveelheden *instant messaging,* chat sessies, sociale netwerken, geluid, foto's en video die op ons afkomen?

Daarvoor zullen we nieuwe technieken en principes moeten gebruiken die wetenschappers, studenten en bedrijven nu ontwikkelen.

We zullen verder door moeten gaan met relatief eenvoudig en dom werk door computerprogramma's te laten uitvoeren. Daar moet hoogwaardiger werk voor terug komen. Dat moeten we dan natuurlijk wel met zijn allen kunnen oppakken. Onderwijs van hoge kwaliteit en toegankelijk voor iedereen, blijft dus noodzakelijk in de toekomst!

Want, mijns inziens zullen mensen altijd noodzakelijk blijven voor kwaliteitscontrole en kwaliteitswaarborging. Ook zullen ze moeten zorgen voor "serendipiteit": nieuwe interesses aanwakkeren en verrassingen tonen, want daar is meer voor nodig dan willekeurige selecties gemaakt door computerprogramma's, zoals dat nu soms gaat.

Ook moeten we blijven begrijpen hoe computerprogramma's werken en wat de beperkingen zijn, zoals u nu weet wat de beperkingen van internet zoekmachines zijn. Ik heb er echter vertrouwen in dat mensen altijd de beperkingen van machines snel genoeg in de gaten zullen krijgen.

Als altijd blijft onze privacy belangrijk, we moeten controle blijven houden over onze informatie. Hoewel dat meer lijkt te gelden voor ouderen dan voor jongeren: de laatsten lijken al meer gewend aan het relatieve gebrek aan privacy of ze laten zich minder snel in de maling nemen.

En die twintig jaar die het duurt voor we allemaal dagelijks één of meer picabytes te verwerken krijgen, die zullen we hard nodig hebben om nieuwe technieken te ontwikkelen, deze toepasbaar te maken, en ze op voldoende data te testen en te verbeteren. Maar we hebben die tijd ook vooral nodig om met zijn allen aan al die nieuwe technieken en nieuwe manieren van werken te wennen!

10 Dankwoord

Ten eerste wil ik het College van Bestuur van de Universiteit van Maastricht hartelijk danken voor hun bereidheid te investeren in de leerstoel text mining. Zoals aangegeven investeert zij hiermee niet alleen in onderwijs en onderzoek in de technologie zelf, maar ook in het toegankelijk blijven van informatie in de toekomst. Volgende maand zal het vak *text mining* als verplicht onderdeel van de Masters opleiding van start gaan, en ik verheug me op de samenwerking met studenten en nieuwe collega's om de eerste versie van de cursus tot een succes te maken!

Ik wil ZyLAB Technologies BV, en mijn collega's in het bijzonder danken voor het financieren van de bijzondere leerstoel.

In deze gaat speciale dank uit naar de collega hoogleraren Jaap van den Herik en Eric Postma die met veel doorzettingsvermogen geholpen hebben deze bijzondere leerstoel te realiseren. Het gehele proces begon een aantal jaren geleden in Delft op initiatief van Jaap, die het vervolgens niet heeft losgelaten. Het is jammer dat jullie Maastricht verlaten hebben en naar Tilburg zijn gegaan, maar zoals ik aangegeven heb, zal ik in Maastricht de honneurs voor jullie waarnemen!

Jaap verdient een extra dankwoord, want onder hem ben ik indertijd mijn wetenschappelijke carrière begonnen, in 1985 aan de TU-Delft wel te verstaan. Jaap: na mijn afstuderen vertelde ik je dat ik een punt zette achter mijn wetenschappelijke carrière. Toen ik later toch besloot om te gaan promoveren aan de Universiteit van Amsterdam, kwam ik ermee

weg door je te vertellen dat het puntkomma was geworden. In het kader van deze laatste stap was het wellicht toch een komma, maar daar heb je zelf dan ook hard aan meegewerkt. Jaap, bedankt voor alles!

Ook gaat dank uit naar Professor Remco Scha, bij wie ik in 1993 promoveerde aan de Universiteit van Amsterdam. Dank voor je flexibiliteit die mij toestond om te promoveren en tegelijk aan mijn eigen bedrijf te werken. Ook dank voor de vele sturing en in het bijzonder voor de tip die me nu nog bij staat: "een promotie is iets anders dan een verkoopverhaal, het was waarschijnlijk de enige keer in mijn leven dat ik ergens echt hard over na kon denken". Je had gelijk, het kan geen kwaad om ergens echt de tijd voor te nemen en af en toe eens heel diep na te denken.

Verder gaat mijn dank uit naar de Koninklijke Marine, en in het bijzonder de Marine Inlichtingen Dienst, voor de *hands-on* training en de unieke ervaring op een bijzonder gebied van de informatie technologie die voor weinigen toegankelijk dan wel bekend is.

In deze context ben ik ook dank verschuldigd aan alle klanten van ZyLAB, in het bijzonder diegenen die hun grootste informatie technische problemen met ons wilden delen en samen met ons aan nieuwe oplossingen wilden werken. Zelfs als dat betekende dat we het risico moesten nemen om nieuwe programma's en nieuwe technieken te ontwikkelen. Door mee te werken aan het oplossen van deze problemen, hebben ZyLAB en ikzelf in het bijzonder altijd vooraan gelopen bij nieuwe inzichten in nieuwe technieken om de meest complexe informatie technische problemen van de laatste twintig jaar op te lossen met innovatieve, maar vooral ook praktische en gebruikersvriendelijke oplossingen. De lijst is waarschijnlijk te lang om op te noemen, maar de UN Oorlogstribunalen, de Europese Commissie en diverse verder niet bij naam te noemen (internationale) opsporings- en inlichtingendiensten verdienen een speciale vermelding. Ik hoop in de toekomst verder met hen te werken bij het ontwikkelen van vele nieuwe toepassingen van informatie technologie.

Mijn familie en ouders hebben altijd een belangrijke rol voor me gespeeld. Ik wil ze hierbij allemaal bedanken dat ze altijd achter me gestaan hebben. Ria, wat zou Kees trots geweest zijn, helaas kan hij er niet bij zijn. Het was mijn vader die mij min of meer inschreef voor de informatica opleiding aan de TU-Delft. Hij was het ook die, vanwege

mijn dreigende militaire dienst als soldaat bij de landmacht, ervoor zorgde dat de Marine Inlichtingen Dienst in het bezit kwam van mijn afstudeerverslag, wat al snel leidde tot een plaatsing als officier bij de Koninklijke Marine in Amsterdam. Als zelfstandig ondernemer heeft hij mijn ZyLAB activiteiten altijd volledig ondersteund, maar ook mijn wetenschappelijke activiteiten heeft hij altijd gestimuleerd.

Tot slot wil ik mijn vrouw Frédérique en mijn kinderen Josefien, Stefanie en Loek ook bedanken voor hun onvoorwaardelijke steun en liefde. Frédérique heeft meer dan wie dan ook dit hoogleraarschap gestimuleerd en ervoor gezorgd dat ik er de nodige tijd en aandacht aan heb kunnen besteden, wat vaak ook ten koste ging van tijd en aandacht voor de familie! Frédérique, Josefien, Stefanie en Loek: vandaag is ook jullie feestje!

Dames en heren, dank voor uw aandacht.

Ik heb gezegd.

11 Verwijzingen en noten

In deze paragraaf zijn per onderdeel noten en verwijzingen naar aanvullende literatuur opgenomen.

Wat is Text Mining?

Scholtes, J.C. (2008d) geeft een kort overzicht over technieken die gebruikt worden in *text analytics* en *text mining*. In Witten, I.H. and Frank, E. (2005) staat een uitgebreid overzicht van het vakgebied *data mining,* de gestructureerde variant van *text mining*.

Andere basis boeken op het gebied van *text mining* worden vermeld in de sectie over *text mining* technieken. De meest toonaangevende op dit moment zijn: Feldman, R., and Sanger, J. (2006), Berry, M.W., Editor (2004) en Berry, M. W. and Castellanos, M. Editors (2006).

Zoeken met Computers in Ongestructureerde Informatie

Blair, D.C. and Maron, M.E. (1985) was het eerste onderzoek dat de effectiviteit van puur Booleaanse zoeksystemen in twijfel trok in 1985. De conclusies worden nog steeds bevestigd. Recent weer door het LEGAL-TREC onderzoek en door Baron, Jason R. (2005).

Andrews, Whit and Knox, Rita (2008) geeft een goed overzicht van commercieel verkrijgbare *Information Access* systemen: systemen die een combinatie van zoeken, visualisatie, *text mining* en integratie met andere business applicaties bieden.

Voor meer informatie over de werking van zoekmachines wordt verwezen naar een groot aantal klassiekers uit de *informatie retrieval* literatuur. Deze publicaties geven een uitgebreid overzicht geven van de diverse zoek-, relevance ranking- en programmeertechnieken die in de loop der jaren ontwikkeld zijn voor het zoeken binnen grote hoeveelheden tekst. Soms puur wiskundige technieken, maar in de loop der jaren ook meer en meer technieken die gebruik maakten van *artificial intelligence* en taaltechnologie: Crestani, F., Lalmas, M. and Rijsbergen, C.J. van, (Editors), 1998, Croft, W.B. and Harper, D.J. (1979), Croft, Bruce (Editor), (2000), Dominich, Sándor (2008), Grefenstette, Gregory (1998), Kowalski, Gerald (1997), Kruschwitz, Udo (2005), Losee, R.M. (1998), Manning, Christopher

D., Raghavan, Prabhakar, and Schütze, Hinrich (2008), Meadow, C.T., Boyce, B.R., Kraft, D.H. and Barry, C. (2007), Rijsbergen, C.J. van (1979), Rijsbergen, C.J. van (2004), Salton, G., Wong, A. and Yang, C.S. (1968), Salton, Gerard (1971), Salton, Gerard, (1975), Salton, Gerard, and McGill, Michael (1983), Salton, Gerard, (1989), Scholtes, J.C. (1995), Scholtes, J.C. (1996), Scholtes, J.C. (2007g), Spink, Amanda and Cole, Charles (Editors), (2005), Tait, John I. (Editor), (2005), White, Martin (2007), en Wilkingson, R., Arnold-Moore, T., Fuller, M., Sacks-Davis, R., Thom, J. and Zobel, J. (1998).

Knuth, D.E. (1998) en Knuth, D.E. (2008) geven een goed overzicht van de onderliggende algoritmes die zoekmachines gebruiken.

Text Mining in Relatie tot "Zoeken & Vinden"

Scholtes, J.C. (2005a) en Scholtes, J.C. (2009) gaan in meer detail in waarom en wanneer het relevant is om "alles" te vinden in plaats van alleen de meest relevante documenten. Ook wordt ingegaan op technieken om zaken te vinden die niet gevonden willen worden en hoe men zaken vindt terwijl men niet precies weet waar men op zoekt.

Ingwersen, Peter and Järvelin, Kalervo (2005) beschrijft diverse technieken om gebruik te maken van de context van een document of kennis over een domein om beter en efficiënter te zoeken.

Over *text mining* en informatie visualisatie is een keur aan literatuur beschikbaar.

Een van de meest toonaangevende en volledigste is Card, Stuart K., Mackinlay, Jock D., and Shneiderman, Ben, Editors (1999), waarin een overzicht wordt gegeven van bijna alle visualisatie technieken die tot 2000 beschikbaar waren. Ook de herdruk van Tufte, Edward, R. (2001) is een absolute aanrader.

Andere referenties zijn: Bimbo, Alberto del (1999), Chen, Chaomei (2006), Fry, Ben (2008), en Scholtes, J.C. (2005b).

Meer over andere voordelen en toepassingen van *text mining* om van ongestructureerde data gestructureerde data te maken zijn te vinden in: Chakrabarti, S. (2003) en in Chan, G., Healey, M.J., McHugh, J.A.M., and Wang, J.T.L., (2001).

Voorbeelden van Toepassingen van Text Mining

In Knox, R. (2008) staat een uitgebreid overzicht van commerciële toepassingen van *text mining,* speciaal gericht op de strategische toepassing van *text mining* binnen een IT-organisatie.

Miller, Thomas W. (2005), Prado, Hercules Antonio Do (Editor), Ferneda, Edilson (Editor), (2008), Spangler, Scott and Kreulen, Jeffrey (2008), en Sullivan, Dan (2001) bevatten meerdere goede beschrijvingen van de praktische en commerciële toepassingen van *text mining* technologie.

Voor de toepassing van *text-mining* binnen fraude- en criminaliteits-opsporing en inlichtingen analyses wordt verwezen naar Scholtes, J.C. (2007a), Scholtes, J.C. (2007b), Scholtes, J.C. (2007c), Scholtes, J.C. (2007d), Scholtes, J.C. (2008b) en natuurlijk DARPA: Defense Advanced Research Project Agency (1991).

Voorbeelden van de toepassing van *text mining* voor *business intelligence* kunnen gevonden worden in: Halliman, Charles (2001) en in Inmon, William H. and Nesavich, Anthony (2008).

Technieken en toepassingen die gebruikt worden bij *sentiment mining* zijn te vinden in Shanahan, J.G., Qu, Y., and Wiebe, J. (Editors), (2006) en in Scholtes, J.C. (2008).

Meer over *text mining* bij klinisch onderzoek en andere biomedische toepassingen is te lezen in Herron, Patrick (2008), Zvelebil, M. and Baum, J.O. (2008) en in Ananiadou, Sophia (Editor), Mcnaught, John (Editor), (2006).

E-discovery is in potentie één van de meest veelbelovende toepassings-gebieden van *text mining*, zeker binnen de context van de kredietcrisis en alle onderzoeken en rechtszaken die gegarandeerd gaan volgen.

Meer over de wet- en regelgeving van *e-discovery* en de *Federal Rules of Civil Procedure* kan gevonden worden in Dahlstrom Legal Publishing (2006), op EDRM (Electronic Discovery Reference model): http://www.edrm.net, in Paul, G.L. and Nearon, B.H. (2006) en in *The Discovery Revolution. E-Discovery Amendments to the Federal Rules of Civil Procedure*. American Bar Association.

Debra Logan, John Bace, and Whit Andrews (2008) geeft een zeer volledig overzicht van commerciële leveranciers van *e-discovery* software oplossingen.

Meer referenties voor advocaten en juristen over *e-discovery* kunnen gevonden worden in Lange, M.C.S. and Nimsger, K.M. (2004), op de Sedona Conference website: http://www.thesedonaconference.org/ en in Socha, George (2009). Dit laatste rapport gaat over het in-huis uitvoeren van delen van het *e-discovery* proces met de bijbehorende risico's en voordelen.

Meer gedetailleerde beschrijvingen van *e-discovery* technieken en toepassingen in relatie tot *text mining* en *information retrieval* kunnen gevonden worden in Scholtes, J.C. (2006c), Scholtes, J.C. (2007f), Scholtes, J.C. (2007h), Scholtes, J.C. (200i7), Scholtes, J.C. (2007j), en Scholtes, J.C. (2008c).

In de komende jaren zal nieuwe regelgeving en *compliance* een belangrijk onderwerp worden. Meer hierover en over de toepassingen van *text mining* en *information retrieval* in relatie tot email, records management en fraude opsporing kan gevonden worden in: Manning, George A. (2000), Scholtes, J.C. (2004a), Scholtes, J.C. (2004b), Scholtes, J.C. (2005c), Scholtes, J.C. (2005d), Scholtes, J.C. (2006a), Scholtes, J.C. (2006b), Scholtes, J.C. (2007e), Scholtes, J.C. (2008f), en Scholtes, J.C. (2007k).

De Technologie achter Text Mining

De core-technologie achter text mining is zeer uitgebreid en gedetailleerd na te lezen in: Feldman, R., and Sanger, J. (2006), Berry, M.W., Editor (2004), Berry, M. W. en Castellanos, M. Editors (2006) en Weiss, et al. (2005). De eerste referentie zal gebruikt worden als tekstboek bij het college.

Een bijzonder boek is Bilisoly, Roger (2008), hierin wordt met behulp van *Perl* het gebruik van reguliere expressies tot het uiterste doorgevoerd. Zeker interessant voor fans van de programmeertaal *Perl*.

Er is veel geschreven over natuurlijk taalverwerking, oftewel *natural language processing* (NLP). Mitkov, Ruslan (2003). *The Oxford Handbook of Computational Linguistics,* is een van de meest complete overzichtswerken. Een van de eerste werken over *discourse analysis* kan gevonden worden

in: Scha, R. and Polanyi, L. (1988). Kay, Martin (1986) en Woods, W.A. (1970) geven meer inzicht in snelle technieken om een grammaticale analyse te maken *(parsing)*. Manning, Christopher D. and Schütze, Hinrich, (1999) is het grote standaardwerk op het gebied van statistische taalverwerking. In Scha, R., Bod, R. and Sima'an, K. (1999) en in Bod, R., Scha, R., and Sima'an, K. (Editors), (2003) wordt het parsen van taal aan de hand van een geanoteerd corpus beschreven: *Data-Oriented Parsing*. En Kao A., Poteet, S. R. (Editors), (2007) beschrijft de rol van natuurlijke taalverwerking binnen t*ext mining* in detail.

Meer over machinaal vertalen kan gevonden worden in: Goutte, C., Cancedda, N., Dymetman, M. and Foster, G. (Eds.). (2009). En tot slot beschrijft Moens, Marie-Francine, (2000) de diverse technieken die beschikbaar zijn voor het automatisch samenvatten van teksten.

Als standaardwerken over patroonherkenning gelden Devijver, P.A. and Kittler, J. (1982), Duda, R.O. and Hart, P.E. (1973) en in de recente bijgewerkte 2e editie: Duda, R.O. and Hart, P.E. (2001). Andere goede bronnen zijn: Bishop, C.M. (2006) en Chen, Y., Li, J., and Wang, J. (2004).

In Moens, Marie-Francine (2006) vinden we een zeer volledig en overzichtelijke uiteenzetting van de bekendste informatie extractie technieken.

Zoals eerder is aangeven, was het de Amerikaanse overheid die een eerste aanzet heeft gegeven voor de extractie van *named entities* uit vrije tekst. Meer hierover kan gevonden worden in een van de weinige openbare publicaties: DARPA: Defense Advanced Research Project Agency (1991).

Sparck-Jones, K. (1971) en Allan, James (Editor), (2002) geven een goed overzicht van de visie van traditionele *information retrieval* specialisten op entiteit-extractie.

Meer over *machine learning* kan gevonden worden in de klassieke werken van Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986a) en Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986a), en in Mitchell, Tom, (1997). *Machine Learning*. McGraw Hill.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005) geeft een goed overzicht van het gebruik van *machine learning* technieken voor tekst classificatie.

Meer over *Support Vector Machines* (SVM) kan gevonden worden in Cristianini, N. and Shawe-Taylor, J. (2000).

Een andere klassieker is de wetenschappelijke publicatie met de titel *Latent Semantic Indexing* van Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwater, S. and Harshman, R. (1988). Ook Voorhees, Ellen M. (1985) is een interessant overzicht van de toepassing van cluster technieken binnen *information retrieval*.

Vervolgens is er veel onderzoek gedaan in het begin van de jaren negentig door ondergetekende naar toepassingen van zelforganiserende neurale netwerken voor taalverwerking en *information retrieval*. Meer kan gevonden worden in: Kohonen, T. (1984) en in Scholtes, J.C. (1990a). Scholtes, J.C. (1990b). Scholtes, J.C. (1991a), Scholtes, J.C. (1991b). Scholtes, J.C. (1991c). Scholtes, J.C. (1991d). Scholtes, J.C. (1991e). Scholtes, J.C. (1991f). Scholtes, J.C. (1991g). Scholtes, J.C. (1991h). Scholtes, J.C. (1991i). Scholtes, J.C. (1991j). Scholtes, J.C. (1991k). Scholtes, J.C. (1992a). Scholtes, J. (1992b). Scholtes, J.C. (1992c). Scholtes, J.C. (1992d). Scholtes, J.C. (1992e). Scholtes, J.C. (1992f). Scholtes, J.C. (1992g). Scholtes, J.C. and Bloembergen, S. (1992a). Scholtes, J.C. and Bloembergen, S. (1992b). Scholtes, J.C. (1992h). Scholtes, J.C. (1993). Scholtes, J.C. (1994a). Scholtes, J.C. (1994b).

Onderwijs en Onderzoek

In Baron, Jason R. (2005) is meer te vinden over de grote voortrekker van het LEGAL-TREC initiatief om zoektechnieken te evalueren zodat ze betrouwbaar in rechtszaken kunnen worden ingezet. Details over het Legal-TREC Research Program zijn hier te vinden: *http://trec-legal.umiacs.umd.edu/*.

Jason Baron is ook betrokken bij de *Sedona Conference,* een initiatief van diverse advocaten, bedrijfsjuristen en rechters om standaarden te definiëren op het gebied van *e-discovery:* Sedona Conference: *http://www.thesedonaconference.org/*.

LEGAL-TREC was een voortzetting van TREC, meer over de geschiedenis, doelstellingen en resultaten van TREC kan hier gevonden worden: Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005).

Konchady Manu, (2006) is een goed praktisch werkboek dat in combinatie met de nodige *open source text mining* software gebruikt gaat worden tijdens het praktische gedeelte van het *text mining* college.

Conclusies en Vooruitblik

Meer over *optical character recognition (OCR)* kan gevonden worden in Henseler, J., Scholtes, J.C., and Verhoest, C.R.J. (1987) en Herik, H.J. van den, Scholtes, J.C. and Verhoest, C.R.J. (1988).

Jurafsky, D. and Martin, J.H., (2009) geeft een breed overzicht over spraakherkenning technologie.

Een intrigerend boek is Kurzweil, Ray (2005). Hierin wordt door de maker van één van de eerste commerciële OCR machines een bijzondere visie gegeven over de gevolgen van de informatie maatschappij en de convergentie van mensen en machines.

Andere boeiende publicaties over de maatschappelijke impact die recente ICT technieken tot gevolg gehad hebben voor massa collaboratie, het oplossen van complexe problemen, het verschijnsel dat ook wel *Wishdom of Crowds* wordt genoemd en het "concurreren door te analyseren" kunnen gevonden worden in: Tapscott, D. and Williams, A.D. (2006), Ayers, Ian (2007), Davenport, T.H. and Harris, J.G. (2007), Segaran, T. (2007), en Surowiecki, James (2004). Al deze nieuwe maatschappelijke en economische principes zijn mogelijk geworden door de toepassing van text mining technieken.

Meer over het zoeken op inhoud in multimediale bestanden kan gevonden worden in: Postma E.O. and Herik, H.J. van den (2000), Wu, J.K., Kankanhalli, M.S., Lim, J.H., and Hong, D. (2000) en in Wang, James Z. (2001).

Uitleg over de architectuur en de algoritmes die gebruikt worden in *The Grid* zijn te vinden in: Berman, F., Fox, G. and Hey, T. (Editors), (2003), Li, Maozhen and Baker, Mark (2005) en Liu, Bing (2007).

Meer over kwantum computers en kwantum algoritmes kan gevonden worden in: Kaye, P., Laflamme, R. and Mosca, M. (2007) en in Steeb, W.H. and Hardy, Y. (2006).

En Scholtes, J.C. (2008e) geeft een korte visie op de toekomst en over de mogelijke risico's en voordelen van de moderne informatie maatschappij.

12 Literatuurlijst

Allan, James (Editor), (2002). *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers.

Ananiadou, Sophia (Editor), Mcnaught, John (Editor), (2006). *Text mining for biology and biomedicine*. Artech House.

Andrews, Whit and Knox, Rita (2008). *Magic Quadrant for Information Access Technology.* September 30, 2008. Gartner Research Report, ID Number: G00161178. Gartner, Inc.

Ayers, Ian (2007). Super Crunchers. *Why Thinking-by-Numbers is the New Way to be Smart*. Bantam Books.

Baron, Jason R. (2005). Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery. S*edona Conference Journal*. Vol. 6, 2005.

Berman, F., Fox, G. and Hey, T. (Editors), (2003). *Grid computing: making the global infrastructure a reality.* John Wiley and Sons.

Berry, M.W., Editor (2004). *Survey of text mining: clustering, classification, and retrieval*. Springer-Verlag.

Berry, M. W. and Castellanos, M. Editors (2006). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer-Verlag.

Bilisoly, Roger (2008). *Practical Text Mining with Perl* (Wiley Series on Methods and Applications in Data Mining). John Wiley and Sons.

Bimbo, Alberto del (1999). *Visual Information Retrieval*. Morgan Kaufmann.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

Blair, D.C. and Maron, M.E. (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM,* Vol. 28, No. 3, pp. 289-299.

Bod, R., Scha, R., and Sima'an, K. (Editors), (2003). *Data-Oriented Parsing*. Center for the Study of Language and Information, Stanford, CA.

Card, Stuart K., Mackinlay, Jock D., and Shneiderman, Ben, Editors (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers.

Chakrabarti, S. (2003). Mining the Web. *Discovering Knowledge from Hypertext Data*. Morgan Kaufman.

Chan, G., Healey, M.J., McHugh, J.A.M., and Wang, J.T.L., (2001). *Mining the World Wide Web, an information search approach*. Kluwer Academic Publishers.

Chen, Chaomei (2006). *Information Visualization: Beyond the Horizon*. Springer-Verlag.

Chen, Y., Li, J., and Wang, J. (2004). *Machine Learning and Statistical Modelling Approaches to Image Retrieval*. Kluwer Academic Publishing.

Crestani, F., Lalmas, M. and Rijsbergen, C.J. van, (Editors), 1998. *Information Retrieval: Uncertainty and Logics. Advanced Models for the Representation and Retrieval of Information*. Kluwer Academic Publishers.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press.

Croft, W.B. and Harper, D.J. (1979). Using Probabilistic Models of Information Retrieval without Relevance Information. *Journal of Documentation*. Vol. 35, No. 4, pp. 285-295.

Croft, Bruce (Editor), (2000). *Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval*. Kluwer Academic Publishers.

Dahlstrom Legal Publishing (2006). The New *E-Discovery Rules. Amendments to the Federal Rules of Civil Procedure Addressing Discovery of Electronically Stored Information (effective December 1st, 2006)*.

DARPA: Defense Advanced Research project Agency (1991). Message Understanding Conference (MUC-3). *Proceedings of the Third Message Understanding Conference (MUC-3)*. DARPA.

Davenport, T.H. and Harris, J.G. (2007). *Competing on Analytics. The New Science of Winning*. Harvard Business School Press.

Devijver, P.A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice Hall.

Dominich, Sándor (2008). *The Modern Algebra of Information Retrieval*. Springer-Verlag.

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. John Wiley and Sons.

Duda, R.O. and Hart, P.E. (2001). *Pattern Classification (2nd Edition)*. John Wiley and Sons.

Dumais, S.T., Furnas, G.W., Landauer, T.K. , Deerwater, S. and Harshman, R. (1988). Using Lantent Semantic Analysis to Improve Access to Textual Information. *ACM CHI'88*. pp. 281-285.

EDRM: Electronic Discovery Reference model: *http://www.EDRM.net*

Escher, M.C. Official M. C. Escher Web site, published by the M.C. Escher Foundation and Cordon Art B.V. *http://www.mcescher.com/*

Feldman, R., and Sanger, J. (2006). The Text Mining Handbook: *Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.*

Fry, Ben (2008). *Visualizing Data. Exploring and Explaining Data with the Processing Environment*. O'Reilly.

Grefenstette, Gregory (1998). *Cross-Language Information Retrieval*. Kluwer Academic Publishers.

Grossman, D.A. and Frieder, O. (2004). Information Retrieval: *Algorithms and Heuristics (The Information Retrieval Series)*. Springer-Verlag.

Goutte, C., Cancedda, N., Dymetman, M. and Foster, G. (Eds.) (2009). *Learning Machine Translation.* MIT Press.

Halliman, Charles (2001). *Business Intelligence Using Smart Techniques. Environmental Scanning Using Text mining and Competitor Analysis Using Scenarios and Manual Simulation.* Information Uncover.

Henseler, J., Scholtes, J.C., and Verhoest, C.R.J. (1987). *The Design of a Parallel Knowledge-Based Optical Character-Recognition System.* M.Sc. Thesis. Delft University of Technology, Department of Mathematics & Computer Science, 1987.

Herik, H.J. van den, Scholtes, J.C. and Verhoest, C.R.J. (1988). The Design of a Knowledge-Based Optical-Character Recognition System. *Proc. of the SCS,* June 1-3, 1988, pp. 350-358. Nice, France.

Herron, Patrick (2008). *Text Mining for Genomics-based Drug Discovery.* VDM Verlag Dr. Müller.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text Classification Using Machine Learning Techniques, *WSEAS Transactions on Computers,* Issue 8, Volume 4, August 2005, pp. 966-974.

Ingwersen, Peter and Järvelin, Kalervo (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer-Verlag.

Inmon, William H. and Nesavich, Anthony (2008). *Tapping into Unstructured Data. Integrating Unstructured Data and Textual Analytics into Business Intelligence*. Prentice Hall.

Jurafsky, D. and Martin, J.H., (2009). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition*. Pearson, Prentice Hall.

Kao A., Poteet, S. R. (Editors), (2007). *Natural Language Processing and Text Mining*. Springer-Verlag.

Kay, Martin (1986). *Algorithm schemata and data structures in syntactic processing. Readings in natural language processing*. Morgan Kaufmann Publishers Inc.

Kaye, P., Laflamme, R. and Mosca, M. (2007). *An Introduction to Quantum Computing*. Oxford Press.

Konchady Manu, (2006). *Text Mining Application Programming (Programming Series)*. Charles River Media.

Kowalski, Gerald (1997). *Information Retrieval Systems*. Theory and Implementation. Kluwer Academic Publishers.

Knox, R. (2008). Content Analytics Supports many Purposes. *Gartner Research Report,* ID Number: G00154705, January 10, 2008.

Knuth, D.E. (1998). *Art of Computer Programming, Volume 1-3 (2nd Edition)*. Addison Wesley Professional.

Knuth, D.E. (2008). *The Art of Computer Programming, Volume 4, Fascicle 0: Introduction to Combinatorial Algorithms and Boolean Functions*. Addison Wesley Professional.

Kruschwitz, Udo (2005). *Intelligent Document Retrieval. Exploiting Markup Structure*. Springer-Verlag.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Springer-Verlag.

Kurzweil, Ray (2005). *The Singularity is Near, when Humans Transcend Biology*. Viking (Penguin Group).

Logan, Debra, Bace, John, and Andrews, Whit (2008). *MarketScope for E-Discovery Software Product Vendors*. Gartner Research Report ID Number: G00163258. Gartner, Inc.

Lange, M.C.S. and Nimsger, K.M. (2004). *Electronic Evidence and Discovery: What Every Lawyer Should Know*. American Bar Association.

Legal-TREC Research Program: *http://trec-legal.umiacs.umd.edu/*.

Li, Maozhen and Baker, Mark (2005). *The Grid: Core Technologies*. John Wiley and Sons.

Liu, Bing (2007). *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag.

Losee, R.M. (1998). *Text-Retrieval and Filtering. Analytic Models of Performance*. Kluwer Academic Publishers.

Manning, Christopher D. and Schütze, Hinrich, (1999). *Foundations of statistical natural language processing*. MIT Press.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, George A. (2000). *Financial Investigation and Forensic Accounting*. CRC Press.

Meadow, C.T., Boyce, B.R., Kraft, D.H. and Barry, C. (2007). *Text Information Retrieval System (3rd Edition)*. Academic Press, Elsevier.

Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986a). *Machine Learning, an Artificial Intelligence Apporach. Volume 1*. Morgan Kaufmann.

Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986b). *Machine Learning, an Artificial Intelligence Apporach. Volume 2*. Morgan Kaufmann.

Miller, Thomas W. (2005). *Data and text mining: a business applications approach*. Pearson Prentice Hall.

Mitchell, Tom, (1997). *Machine Learning*. McGraw Hill.

Mitkov, Ruslan (2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Moens, Marie-Francine, (2000). *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers.

Moens, Marie-Francine (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag.

Paul, G.L. and Nearon, B.H. (2006). *The Discovery Revolution. E-Discovery Amendments to the Federal Rules of Civil Procedure*. American Bar Associaton.

Postma E.O. and Herik, H.J. van den (2000). Discovering the visual signature of painters. In N. Kasabov (Editor), *Future Directions for Intelligent Systems and Information Sciences. The Future of Speech and Image Technologies, Brain Computers, WWW and Bioinformatics, pp. 129-147.* Heidelberg: Physica Verlag (Springer Verlag).

Prado, Hercules Antonio Do (Editor), Ferneda, Edilson (Editor), (2008). *Emerging technologies of text mining: techniques and applications.* Information Science Reference.

Rijsbergen, C.J. van (1979). *Information Retrieval.* Butterworths, London.

Rijsbergen, C.J. van (2004), *The Geometry of Information Retrieval.* Cambridge University Press.

Salton, G., Wong, A. and Yang, C.S. (1968). A Vector Space Model for Automatic Indexing. *Communications of the ACM.* Vol. 18, No. 11, pp. 613-620.

Salton, Gerard (1971). *The Smart Retrieval System.* Prentice Hall.

Salton, Gerard, (1975). *Dynamic information and library processing.* Prentice Hall.

Salton, Gerard, and McGill, Michael (1983). *Introduction to modern information retrieval.* McGraw-Hill.

Salton, Gerard, (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley.

Scha, R. and Polanyi, L. (1988). An Augmented Context Free Grammar for Discourse. *Proceedings of the 12th International Conference on Computational Linguistics.* Budapest, August 1988, pp. 573-577.

Scha, R., Bod, R. and Sima'an, K. (1999). A Memory-Based Model of Syntactic Analysis: Data-Oriented Parsing. *Journal of Experimental and Theoretical Artificial Intelligence (Special Issue on Memory-Based Language Processing, edited by Walter Daelemans).* Vol. 11, Nr. 3 (July 1999), pp. 409-440.
Scholtes, J.C. (1990a). Trends in Neurolinguistics. *Proceedings of the IEEE Symposium on Neural Networks,* June 21, Delft, Netherlands, pp. 69-86.

Scholtes, J.C. (1990b). *Neurolinguistics*. Computational Linguistics Project, CERVED S.p.A., Italy, 1990.

Scholtes, J.C. (1991a). Recurrent Kohonen Self-Organization in Natural Language Processing. *Artificial Neural Networks* (T. Kohonen, K. Makisara, O. Simula and J. Kangas, Eds.), pp. 1751-1754. Elsevier Science Publishers, Amsterdam, The Netherlands.

Scholtes, J.C. (1991b). Using Extended Kohonen-Feature Maps in a Language Acquisition Model. *Proceedings of the 2nd Australian Conference on Neural Nets*. February 2-4. Sydney, Australia, pp. 38-43.

Scholtes, J.C. (1991c). Learning Simple Semantics by Self-Organization. *Worknotes of the AAAI Spring Symposium Series on Machine Learning of Natural Language and Ontology*. March 26-29. Palo Alto, CA, pp. 146-151.

Scholtes, J.C. (1991d). Learning Simple Semantics by Self-Organization. *Worknotes of the AAAI Spring Symposium on Connectionist Natural Language Processing*. March 26-29. Palo Alto, CA, pp. 78-83.

Scholtes, J.C. (1991e). Unsupervised Context Learning in Natural Language Processing. *Proceedings of the International Joint Conference on Neural Networks*. July 8-12. Seattle, WA, Vol. 1, pp. 107-112.

Scholtes, J.C. (1991f). Self-Organized Language Learning. *The Annual Conference on Cybernetics: Its Evolution and Its Praxis*. July 17-21. Amherst, MA.

Scholtes, J.C. (1991g). Unsupervised Learning and the Information Retrieval Problem. *Proceedings of the International Joint Conference on Neural Networks,* November 18-21, Singapore., pp. 18-21,

Scholtes, J.C. (1991h). Filtering the Pravda with a Self-Organizing Neural Net. *Working Notes of the Bellcore Workshop on High Performance Information Filtering*. November 5-7, Chester, NJ.

Scholtes, J.C. (1991i). Kohonen Feature Maps and Full-Text Data Bases: A Case Study of the 1987 Pravda. *Proceedings of Informatiewetenschap 1991*. December 18. Nijmegen, The Netherlands, pp. 203-220. STINFON.

Scholtes, J.C. (1991j). Kohonen's Self-Organizing Map in Natural Language Processing. *Proceedings of the SNN Symposium.* May 1-2. Nijmegen, The Netherlands, p. 64.

Scholtes, J.C. (1991k). Kohonen's Self-Organizing Map Applied Towards Natural Language Processing. *Proceedings of the CUNY 1991 Conference on Sentence Processing*. May 12-14. Rochester, NY, p. 10.

Scholtes, J.C. (1992a). Neural Data Oriented Parsing. *Proceedings of the 2nd SNN.* April 14-15, Nijmegen, The Netherlands, p. 86.

Scholtes, J. (1992b). Neural Data Oriented Parsing. *Proceedings of the 3rd Twente Workshop on Language Technology.* May 12-13, Twente, The Netherlands.

Scholtes, J.C. (1992c). Filtering the Pravda with a Self-Organizing Neural Net. *Proceedings of the First SHOE Workshop.* February 27-28, Tilburg, The Netherlands, pp. 267-277.

Scholtes, J.C. (1992d). Resolving Linguistic Ambiguities with a Neural Data-Oriented Parsing (DOP) System. *Proceedings of the First SHOE Workshop,* February 27-28, Tilburg, The Netherlands, pp. 279-282.

Scholtes, J.C. (1992e). Resolving Linguistic Ambiguities with a Neural Data-Oriented Parsing (DOP) System. *Artificial Neural Networks 2* (I. Aleksander and J. Taylor, Eds.). Vol. 2, pp. 1347-1350. Elsevier Science Publishers, Amsterdam, The Netherlands.

Scholtes, J.C. (1992f). Neural Nets for Free-Text Information Filtering. *Proceedings of the 3rd Australian Conference on Neural Nets.* February 3-5, Canberra, Australia.

Scholtes, J.C. (1992g). Filtering the Pravda with a Self-Organizing Neural Net. *Proceedings of the Symposium on Document Analysis and Information Retrieval*, March 16-18, 1992, Las Vegas, NV, pp. 151-161.

Scholtes, J.C. and Bloembergen, S. (1992a). The Design of a Neural Data-Oriented Parsing (DOP) Model. *Proceedings of the International Joint Conference on Neural Networks,* June 7-10, Baltimore, MD.

Scholtes, J.C. and Bloembergen, S. (1992b). Corpus Based Parsing with a Self-Organizing Neural Net. *Proceedings of the International Joint Conference on Neural Networks,* November 3-5, Beijing, P.R. China.

Scholtes, J.C. (1992h). Neural Nets versus Statistics in Information Retrieval. A Case Study of the 1987 Pravda. Proceedings of the SPIE Conference on Applications of Artificial Neural Networks III, April 20-24, Orlando, FL.

Scholtes, J.C. (1993). *Neural Networks in Natural Language Processing and Information Retrieval*. Ph.D. Thesis, January 1993, University of Amsterdam, Department of Computational Linguistics, Amsterdam, The Netherlands.

Scholtes, J.C. (1994a). *Neural Networks in Information Retrieval in a Libraries Context*. State-of-the-Art report, PROLIB/ANN, DG XIII, European Commission, Luxembourgh.

Scholtes, J.C. (1994b). *Neural Networks in Information Retrieval in a Libraries Context*. Final report, PROLIB/ANN, DG XIII, European Commission, Luxembourgh.

Scholtes, J.C. (1995). *Report on knowledge and experience sharing for the International Fund for Agriculteral Development* (United Nations), Rome, May 1995.

Scholtes, J.C. (1996). *New Developments in Full-Text Retrieval.* Document 96. Birmingham, September 1996.

Scholtes, J.C. (2004a). What is the single most challenging Sarbanes-Oxley issue today? Ongoing vigilance. *Sarbanes-Oxley Compliance Journal.*

Scholtes, J.C. (2004b). XML for Archiving and Record Management. *The 25th Global Conference and Exhibit*. October 24-28, 2004, Dallas, TX, USA.

Scholtes, J.C. (2005a). Usability versus Precision & Recall. What to do when users prefer a high level of user interaction and ease-of-use over high-tech precision and recall tools. *Search Engine Meeting,* Boston, April 11-12, 2005.
Scholtes, J.C. (2005b). How end-users combine high-recall search tools with visualization. *Intelligence Tools: Data Mining & Visualization, Philadelphia,* June 27-28, 2005.

Scholtes, J.C. (2005c). Affordability in Content Management and Compliance. Knowledge Management World. Best Practices in Enterprise Content Management, May 2005.

Scholtes, J.C. (2005d). From Records Management to *Knowledge Management. Knowledge Management World. Best Practices in Records Management & Regularity Compliance.*

Scholtes, J.C. (2006a). Searching large E-mail collections: the next challenge. *The International Conference for Science & Business Information*, ICIC, Nimes, France. 22-25 October, 2006.

Scholtes, J.C. (2006b). A View on e-mail management. Balancing Multiple Interests and Realities of the Workplace. *Knowledge Management World. Best Practices in E-mail*, February 2006.

Scholtes, J.C. (2006c). Comprehensive eDiscovery and eDisclosure technologies. Next generation deployment of enterprise search tools. *Knowledge Management World. Best Practices in Enterprise Search,* April 2006.

Scholtes, J.C. (2007a). Finding Fraud before it finds you: Advanced Text Mining and other ICT techniques. *Fraud Europe 2007, Brussels, April 24*, 2007.

Scholtes, J.C. (2007b). E-Discovery and e-Disclosure for Fraud Detection. *Fraud World 2007, London, September, 2007.*

Scholtes, J.C. (2007c). Advanced eDiscovery and eDisclosure techniques. *Documation, The Olympia, London, October 2007.*

Scholtes, J.C. (2007d). Roundtable discussion, eDiscovery. David Cooper, J.C. Scholtes, Barry Murphy and Judith Lamonth. *Knowledge Management World,* September 2007.

Scholtes, J.C. (2007e). Efficient and Cost-Effective Email Management with XML. *Knowledge Management World, Best Practices in Email and IM Management*. February 2007.

Scholtes, J.C. (2007f). Mandated e-Discovery Requirement. Comliance Requires Optimal Email Management and Storage. *Today Magazine, the journal of Work Process Improvement*. March/April 2007. pp. 37.

Scholtes, J.C. (2007g). The Evolution of Enterprise Search. *Knowledge Management World. Best Practices in Enterprise Search,* May 2007.

Scholtes, J.C. (2007h). How to make eDiscovery and eDisclosure easier. *AIIM e-Doc Magazine*. Volume 21, Issue 4. July/August 2007. pp. 24-26.

Scholtes, J.C. (2007i). Where Records Management meets Enterprise Search and Knowledge Management: the bundle that optimizes discovery capabilities and supports profitability. *Knowledge Management World. Best Practices in Enterprise Search,* October 2007.

Scholtes, J.C. (2007j). Legal Ease. eDiscovery and eDisclosure. *DM Magazine UK*. November December 2006. pp, 26.

Scholtes, J.C. (2007k). Efficient and Cost-effective *Email Management* With XML. Email Management. (Ms.E jyothi and Elizabeth Raju Eds). Institute of Chartered Financial Analysts of India (ICFAI) Books.

Scholtes, J.C. (2008a). Is there a role for Sentiment Mining in Robot-Human Communications? *First International Conference on Human-Robots Personal Relationships.* June 12-13, 2008.

Scholtes, J.C. (2008b). Finding More: Advanced Search and Text Analytics for Fraud Investigations. *London Fraud Forum, Barbican, London*. October 1, 2008.

Scholtes, J.C. (2008c). Maintain Control During eDiscovery. *Knowledge Management World. Best Practices in eDiscovery,* February 2008

Scholtes, J.C. (2008d). Text Analytics—Essential Components for High-Performance Enterprise Search. *Knowledge Management World. Best Practices in Enterprise Search,* May 2008.

Scholtes, J.C. (2008e). De Hypothese, kolom in Vrij Nederland. 26 Augustus 2008.

Scholtes, J.C. (2008f). Records Management and e-Discovery: Why we need to re-learn Art of Information Destruction. *Knowledge Management World. Best Practices in Records Management and Compliance.* November 2008.

Scholtes, J.C. (2009). Understanding the difference between legal search and Web search: What you should know about search tools you use for e-discovery. *Knowledge Management World. Best Practices in e-Discovery.* January, 2009.

Sedona Conference: *http://www.thesedonaconference.org/*.

Segaran, T. (2007). *Programming Collective Intelligence, Building Smart Web 2.0 Applications*. O'Reilly.

Shanahan, J.G., Qu, Y., and Wiebe, J. (Editors), (2006). *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series).* Springer-Verlag.

Socha, George (2009). *What does it take to bring e-Discovery in-house: risks and rewards.* Published at *www.EDRM.org*.

Spangler, Scott and Kreulen, Jeffrey (2008). *Mining the talk: unlocking the business value in unstructured information.* IBM Press/Pearson plc.

Sparck-Jones, K. (1971). *Automatic Keyword Classification for Information Retrieval.* Butterworths.

Spink, Amanda and Cole, Charles (Editors), (2005). *New Directions in Cognitive Information Retrieval*. Springer Verlag.

Steeb, W.H. and Hardy, Y. (2006). *Problems and Solutions in Quantum Computing and Quantum Information, 2nd edition.* World Scientific Publishers.

Sullivan, Dan (2001). *Document warehousing and text mining*. John Wiley and Sons.

Surowiecki, James (2004). *The Wisdom of Crowds.* Anchor Books.
Tait, John I. (Editor), (2005). *Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Hobour of Karen Spärck Jones.* Springer-Verlag.

Tapscott, D. and Williams, A.D. (2006). *Wikinomics. How Mass Collaboration Changes Everything*. Portfolio (Penguin Group).

Tufte, Edward, R. (2001). *The Visual Display of Quantitative Information, 2nd edition*. Graphics Press.

Voorhees, Ellen M. (1985). The Cluster Hyphothesis Revisited. *Proceedings of the 8th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval*. June 1985, pp. 188-196.

Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005). *TREC: experiment and evaluation in information retrieval*. MIT Press.

Wang, James Z. (2001). Integrated *Region-Based Image Retrieval*. Kluwer Academic Publishers.

Weiss, et al. (2005). Sholom Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau. *Text mining: predictive methods for analyzing unstructured information*. Springer-Verlag.

White, Martin (2007). *Making Search Work. Implementing Web, Intranet and Enterprise Search*. Information Today, Inc.

Wilkingson, R., Arnold-Moore, T., Fuller, M., Sacks-Davis, R., Thom, J. and Zobel, J. (1998). *Document Computing. Technologies for Managing Electronic Document Collections*. Kluwer Academic Publishers.

Witten, I.H. and Frank, E. (2005). *Data Mining, Practical Machine Learning Tools and Techniques, 2nd. Edition*. Morgan Kaufman.

Woods, W.A. (1970). Transition Network Grammars for Natural Language Analysis. *Communications of the ACM*. Vol. 3, Nr. 10, pp. 591-606.

Wu, J.K., Kankanhalli, M.S., Lim, J.H., and Hong, D. (2000). *Perspectives on Content-Based Multimedia Systems*. Kluwer Academic Publishers.

Zvelebil, M. and Baum, J.O. (2008). *Understanding Bioinformatics*. Garland Science, Taylor and Francis Group LLC.

13 English Summary

This text is an extended version of my acceptance speech for the special chair of text mining at the department of knowledge engineering of the University of Maastricht, the Netherlands, presented Friday January 23rd, 2009 at 16.30 hours.

Text-mining refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining encompasses several computer science disciplines with a strong orientation towards artificial intelligence in general, including but not limited to pattern recognition, neural networks, natural language processing, information retrieval and machine learning. An important difference with search is that search requires a user to know what he or she is looking for while text mining attempts to discover information in a pattern that is not known beforehand.

Text mining is particularly interesting in areas where users have to discover new information. This is the case, for example, in criminal investigations, legal discovery and due diligence investigations. Such investigations require 100% recall, i.e., users can not afford to miss any relevant information. In contrast, a user searching the internet for background information using a standard search engine simply requires any information (as oppose to all information) as long as it is reliable. In a due diligence, a lawyer certainly wants to find all possible liabilities and is not interested in finding only the obvious ones.

Increasing recall almost certainly will decrease precision implicating that users have to browse large collections of documents that that may or may not be relevant. Standard approaches use language technology to increase precision but when text collections are not in one language, are not domain specific and or contain variable size and type documents either these methods fail or are so sophisticated that the user does not comprehend what is happening and loses control. A different approach is to combine standard relevance ranking with adaptive filtering and interactive visualization that is based on features (i.e. meta-data elements) that have been extracted earlier.

Other useful applications of text mining can be found in life sciences, compliance, consumer opinions on the internet and product guarantee analysis.

The extra-ordinary chair on text mining of the department of knowledge engineering of the University of Maastricht will focus on text mining methods for language-independent feature extraction for labeling text documents so that large result lists can be filtered and visualized in a meaningful way. Education and research will focus on the application of existing and development of new document classification techniques supporting multi-dimensional and sometimes hierarchical (taxonomy-based) classifications. Meaningful document classification will enable users to dynamically visualize and filter results so they can iteratively refine their search queries.